The horizontal and vertical evolution of Staphylococcus aureus in relation to gene function

Cooper, Jessica E.

Award date:
2005

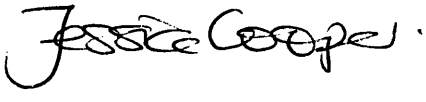
Awarding institution:
University of Bath

THE HORIZONTAL AND VERTICAL EVOLUTION OF *STAPHYLOCOCCUS AUREUS* IN RELATION TO GENE FUNCTION

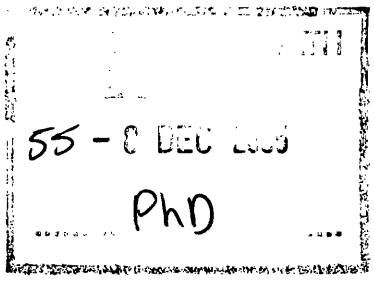
Submitted by Jessica E. Cooper for the degree of Ph.D

of the University of Bath

July, 2005



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346


55 - 8 DEC
PhD

ABSTRACT

This thesis consists of several studies investigating the evolutionary history of the opportunistic human pathogen *Staphylococcus aureus* in relation to gene function. The studies presented are firstly introduced with background to the accumulation of variation in bacterial genomes, the confounding effects of horizontal gene transfer and the clinical relevance of *S. aureus*. Housekeeping genes are typically used for multilocus sequence typing (MLST) and there is little consensus over gene choice for phylogenetic study within, and between, bacterial species at varying phylogenetic levels. The first study compares samples of 'core' genes typical for MLST and evolutionary analyses, which represent different functional categories. These categories are evaluated in terms of functional constraint and the incidence of homologous recombination. Different categories of genes are also assessed for phylogenetic reliability and their utility in the reconstruction of robust phylogeny in this species. The distribution of the 'accessory' surface-protein locus *sdrE* within the natural population of disease and carriage isolates of *S. aureus* is also determined and the resolution provided by the reconstructed phylogeny facilitates the identification of both homologous recombination and gene transfer at this locus. The functional implications of nonsynonymous diversity at this locus are then further considered in the context of Staphylococcal pathogenesis and disease potential. Finally, phylogenetic incongruence within an MLST locus is investigated in relation to the mode of sequence evolution of a further surface protein of close proximity.

I dedicate this thesis to all members of my family, the oldies and the newer additions, for love support and friendship over many years.

xxxxxxxx

ACKNOWLEDGEMENTS

I firstly want to thank my supervisor Ed Feil for his seemingly endless patience when subjected to daily updates of work and personal dramas. Ed has always been supportive and approachable and provided great opportunities. He has become a great friend and I will miss him lots. I must also thank Vicki Fleming. She has been an immensely supportive friend and a great shoulder to lean on through much emotional turmoil! I greatly admire her perseverance and wish her all the best in completing her own study. I am also immensely grateful to Mark Enright who has been a good friend for many years. His encouragement, persuasion and confidence in me got me here in the first place.

I wish to also acknowledge colleagues and collaborators who have helped me in many ways: To Paul Wilkinson for processing endless plates of sequencing and to Sandra Barns for all her help in running the laboratory. Many thanks to Hajo Grundmann for Nottingham isolates and to Sharon Peacock and Nick Day for helpful discussion regarding sdrE. I am most grateful to Timothy Foster for the invitation to Trinity College, Dublin, to conduct platelet aggregation experiments and to Louise O'Brien for being an excellent teacher and great company. Thanks also to Mary Meehan for communications regarding sdrE and to Jean van den Elsen for discussion and the modeling of the sdrE protein. Many thanks also to Eduardo Rocha for much needed assistance with baffling computer programs and the unforgettable Andrew Morgan who conducted preliminary investigations on variation in the region of *clfB*.

I must also thank the beautiful Barbara (Carla) for being a loyal and constant friend through some pretty tough times! Love ya dude! X. Thanks to Jason for support and encouragement (and for his laptop)! Huge love and respect to all friends in Biology and Biochemistry who have made working here such a riot! Special thanks to all the cake bakers whose creations are always infinitely better than my own and to Hinton Garage Rescue for regular roadside recovery!!

CONTENTS

CONTENTS

CONTENTS

CONTENTS

CONTENTS

CHAPTER ONE

BACTERIAL GENE EVOLUTION AND THE BIOLOGY OF *STAPHYLOCOCCUS AUREUS*

CHAPTER ONE: BACTERIAL GENE EVOLUTION AND THE BIOLOGY OF *STAPHYLOCOCCUS AUREUS*

1. INTRODUCTION

The microbial world exemplifies extraordinary diversity. Bacteria are abundant in all corners of our planet, from boiling hot geysers in Yellowstone National Park, to frozen lake water in Antarctica. Bacterial lifestyles can vary from free-living existence in soil and aquatic habitats to symbiotic and parasitic associations in a plethora of plant and animal hosts. Morphology and metabolic capabilities also vary greatly between bacterial taxa. They have an extremely short generation time in the laboratory (typically between 20 minutes and 20 hours dependent upon the species and environmental conditions during lag phase) and large population size facilitating rapid adaptation in response to environmental pressures. These factors and the presence of, typically, only one chromosome makes bacterial populations ideal for the study of evolutionary biology. The adaptive power of microbes has resulted in the emergence and rapid spread of multiple antibiotic-resistance within a matter of decades. The crossing over of genetic material in sexual reproduction is an important mechanism for the generation of diversity in eukaryotic populations. Yet bacteria have an asexual mode of reproduction whereby binary fission produces genetic replica daughter cells. However, the exchange of genetic material in bacterial populations, which seems more likely to introduce new traits rather than by mutation alone, can occur by other mechanisms independent of reproduction. The outcome, however, can be equated with eukaryotic sex since these processes involve the hybridisation of genetic material from separate sources within a single individual (Levin, 1988). However, the species barriers which apply to genetic exchange in eukaryotic populations are essentially absent in bacterial populations complicating aspects of bacterial evolutionary history such as taxonomy and phylogeny and these topics shall be discussed further.

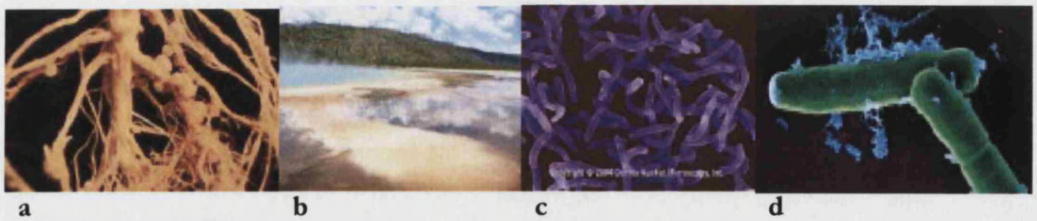


Figure 1. Examples of bacterial diversity. a) Nitrogen-fixing *Rhizobium sp.* can form symbiotic associations with leguminous plants providing the host with organic nitrogen. b) Colourful bacterial mats in hot springs of Yellowstone National Park. c) *Vibrio cholerae* is

the estuarine dwelling causative agent of cholera, one of the biggest killers in the developing world. d) *Bacillus anthracis* is responsible for anthrax, an acute infectious disease which is usually fatal if contracted by inhalation, and is associated with bio-terrorism.

1.1 The accumulation of variation in bacterial populations

The study of the evolutionary history of a population requires an understanding of the mechanisms by the observed genetic variation has arisen.

Mutation

Errors in DNA replication, exposure to gamma radiation, heat and even the action of mobile elements such as transposons can produce local nucleotide change. There are many types of mutation; most involve the substitution of one base for another in a DNA sequence. Single base substitutions (point mutations) can be classified into two categories. A transition is a change of a purine to a purine (A or G) or a pyrimidine to a pyrimidine (C and T). A transversion is a change from a purine to a pyrimidine or vice versa. A single amino acid may be encoded by several different codons which usually differ in the third base of the triplet. This degeneracy of the genetic code means that some substitutions do not change the amino acid sequence of the encoded protein. Such substitutions are termed silent or synonymous. However, some substitutions do have an effect at the protein level. This is termed a missense or nonsynonymous substitution. Indels are another class of mutation which describe insertions and deletions of nucleotides. A single base deletion may result in a frameshift which may subsequently result in a 'stop' codon and the early termination of translation and a truncated protein. This is called a nonsense mutation. Larger sequence deletion can also occur via slipped strand misparing as replication bypasses a folded DNA structure of repeat regions. Breaks in replication can also result in the duplication of part of the chromosome and genome rearrangements.

Gene transfer

New genetic information can be transferred either via homologous recombination whereby homologous sequence to an existing gene is incorporated into the genome replacing the existing sequence. Alternatively, novel genes can be acquired and

incorporated into the genome, this is known as horizontal gene transfer (HGT). HGT is likely to be the mechanism by which organisms acquire the ability to adapt to new ecological niches as observed with the cag pathogenicity island in Helicobacter pylori (Terry, 2005). Homologous recombination introduces variation into already existing genes, rather than introducing unique traits. For successful genetic transfer to occur and an organism to acquire new traits it requires:

- Means for the delivery of genetic material.
- Incorporation of the acquired sequence into the recipient's genome (or plasmid).
- Expression of the incorporated genes in a manner that benefits the organism.

Through the physical contact of donor and recipient cells, as mating pairs or via conjugative pili, bacteria are able to exchange genetic information. This process is called conjugation. Conjugation typically occurs via a self-transmissible or mobilisable plasmid. It can also mediate the transfer of chromosomal sequences by plasmids that integrate into the chromosome and by conjugative transposons, which encode proteins required for their excision from the donor, the formation of a conjugative bridge and transposition into the recipient strain.

Genetic transformation is the uptake of naked DNA from the environment. Unlike conjugation this mechanism allows the acquisition of DNA in the absence of the live donor. Within bacterial species there are varying levels of transformation. Species such as *Neisseria gonnorhoeae* and *Haemophilus influenzae* are constitutively competent. *Bacillus subtilis* and *Streptococcus pneumoniae* are only competent at certain physiological stages in their life cycle. A piece of genetic material is inserted into the chromosome through the incorporation of a single strand to form a stretch of heteroduplex DNA (Smith *et al.*, 1981; Stewart & Carlson, 1986).

For non-transformable species such as *Staphylococcus aureus*, phage-transduction is a very important mechanism for acquiring genetic material. Not all bacteriophage are capable of carrying out transduction and the spectrum of microorganisms that are transducible depends upon receptors recognised by the bacteriophage. The key step in generalised

transduction is the packaging of DNA into phage heads during lytic growth of the phage and bacteriophage and packaged random DNA fragments replicated within a donor microorganism. Some phage enter a lysogenic relationship with its host whereby the lytic cycle is repressed and phage DNA is integrated into the host chromosome at a specific phage attachment site. Upon re-entering the lytic cycle the phage DNA is excised from the host chromosome and may incorporate adjacent host DNA. All phage progeny will then contain this bacterial DNA. This is specialised transduction. The amount of DNA that can be transferred in a single event is limited by the size of the phage capsid, but can range up to approximately 100 kb.

The significance of these mechanisms is that they are not necessarily confined to members of the same species, thus providing a route for the exchange of genetic information over wide taxonomic boundaries. The impact of this will be discussed further in this chapter.

1.2 The fate of polymorphism in bacterial populations

A population in which genotype and allele frequencies remain unchanged over successive generations is said to be in Hardy-Weinberg equilibrium. This model, which usually refers to diploid populations, is based upon five basic assumptions:

1) the population is large resulting in reduced genetic drift (see below)
2) there is no gene flow between populations
3) mutation rates are negligible
4) individuals are mating randomly and
5) natural selection is not operating in the population.

There are no observed examples of natural bacterial populations in Hardy–Weinberg equilibrium. Thus, one or more of the above assumptions is violated and the populations are evolving. For a genotype or an allele to become significant in shaping the evolutionary history of a population it must increase in frequency and ultimately become fixed within the population. The major factors in the rise to frequency are random genetic drift and natural selection.

Random genetic drift

Changes in allele frequency can occur by chance as a result of stochastic forces. This effect is most easily observed in the absence of natural selection (selective neutrality). When all genotypes in the population have the same reproductive fitness, the genotype and allele frequencies may be expected to remain constant over time (Hardy-Weinberg equilibrium). However, random sampling can cause the gene frequencies to change. By chance one genotype may be more reproductively successful, increasing its frequency in the population. Such an increase is random since this genotype is equally likely to be less reproductively successful than other genotypes. Random sampling effects will be greater in smaller populations as the sample is more likely to be biased away from the average. Thus effective population size (defined as the size of an idealized population where the effect of random sampling on allele frequencies would be the same as the actual population) will contribute to the effectiveness of random genetic drift between different populations. However, genetic drift can also occur as a result of fluctuations in selection intensities (Gillespie, 1991). Ultimately, genetic drift may result in the fixation of one allele and the loss of others. Polymorphism will be maintained through the input of alleles into the population by mutation or migration, or by balancing selection.

Natural Selection

Natural selection was proposed by Charles Darwin (1859) to explain phenotypic variation within a species as a result of adaptation. However, Darwin lacked a theory of heredity which was later provided by the rediscovery of Gregor Mendel's ideas in the early 1900's. In the 1930's R. A. Fisher, J. B. S. Haldane and S. Wright demonstrated that Darwin's natural selection and Mendelian heredity are compatible and neo-Darwinism (or the synthetic theory of evolution) is now widely accepted within all areas of biology. Natural selection refers to the differential reproduction of individuals or genotypes within a population resulting from mortality, fertility, fecundicity, reproductive success and the viability of offspring. Most new mutations arising in the population will be deleterious; reducing fitness and will eventually be removed from the population by natural selection. This type of selection is called purifying or negative selection. Should a mutation have no effect on the relative fitness of its carrier it is considered neutral and will be unaffected by selection. In rare cases where a mutation is beneficial to its carrier, conferring a fitness

advantage over other alleles, it will be subjected to positive selection. If a mutation is selectively advantageous it may sweep through the population at the expense of the other variants until it achieves fixation (selective sweep). A low rate of recombination may result in nucleotides linked to this favoured variant hitchhiking with it to also be swept to fixation (hitchhiking effect). Such a process can reduce allelic diversity within the population but not between species (although the concept is problematic in bacteria). Thus the probability of allelic fixation depends upon its frequency, the effective population size and its selective advantage or disadvantage.

1.3 Reassessment of the bacterial clonal paradigm from sequence data

Early studies from multilocus enzyme electrophoresis (MLEE) gave rise to the view that bacterial populations typically exist as a number of discreet clonal lineages. A negligible rate of recombination results in a state of linkage disequilibrium where loci are tightly linked (Levin, 1981). In the presence of extensive recombination, the frequency of a particular allele is independent of the presence of alleles at other loci and the population would be described as in linkage equilibrium. Estimates of linkage disequilibrium have therefore been used to infer rates of recombination, although such methods are fairly insensitive. It is estimated that an allele must change at least 20 times more frequently by recombination than by point mutation in order for the elimination of linkage disequilibrium within a bacterial population (Hudson & Kaplan, 1985; Smith *et al.*, 1993). Therefore even when recombination is more frequent than point mutation there is still significant linkage disequilibrium and only in cases where rates of recombination are extremely high does the population approach linkage equilibrium. Alternatively, recombination could be frequent, but the rise to frequency of a single genotype favoured by selection (transient adaptive clone) would generate disequilibrium. Eventually over time recombination could once again randomise the genetic background. With this in mind it is important to sample the diversity of any population. In particular, sampling bias of disease-associated strains for medically-important species can result in biased estimates of linkage and recombination. Interspecies comparisons can also be problematic and depend upon the sample and number of loci analysed.

Multilocus sequence typing

The development and increased availability of automated sequencing has seen displacement of more laborious typing methods such as pulse field gel electrophoresis (PFGE) and multilocus enzyme electrophoresis (MLEE) with direct nucleotide characterisation schemes. Such nucleotide based typing schemes not only enable the characterisation of strains but nucleotide data can be used to infer evolutionary relationships between strains. Multilocus sequence typing (MLST) was developed on the principle of MLEE which indexes variation based on the electrophoretic mobilities of multiple housekeeping proteins (Maiden *et al.*, 1998). Housekeeping genes are considered to represent the 'core' of the bacterial genome as they encode essential metabolic enzymes. These are unlikely to be subject to selective pressures other than a purifying selection since variation within these genes is likely to be neutral or deleterious. In this way variation is expected to accumulate slowly over time allowing both short and longer term surveillance of bacterial clones. MLST uses the nucleotide sequence of internal fragments of housekeeping genes to directly index variation at the nucleotide level. The higher discrimination afforded by nucleotide sequencing means that similar levels of discrimination are attained by fewer loci than are typically used for MLEE (Maiden *et al.*, 1998). Each unique sequence for a given locus is assigned an allele number. The sequence of allelic integers for a given strain is the sequence type (ST). This process is illustrated in Figure 2.

MLST confers an advantage over single-locus typing (SLST) since schemes are designed with multiple markers scattered around the genome. This greatly increases the discriminatory power of the scheme since they will be robust to the effects of homologous recombination within single loci.

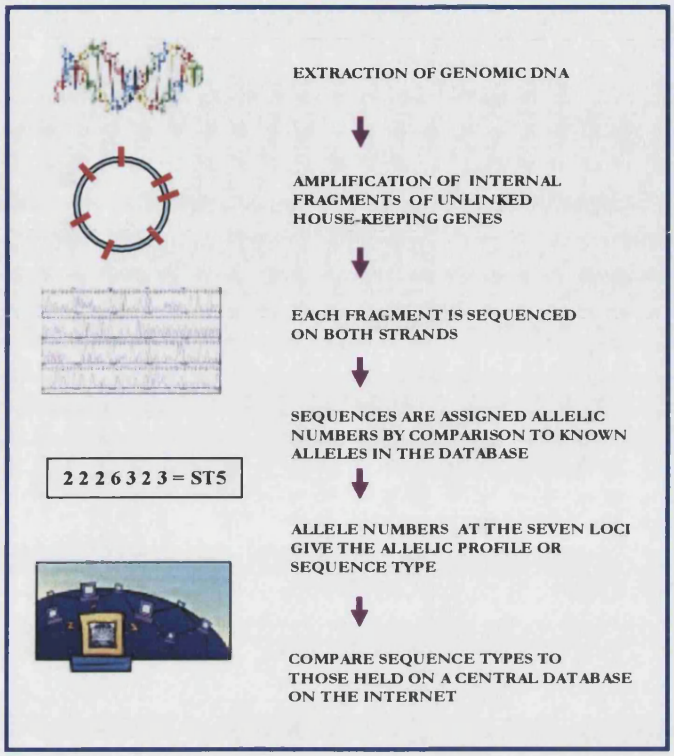


Figure 2. Standard protocol for multilocus sequence typing.

Evolutionary analysis of bacterial species using MLST data

The identification of short term evolutionary relationships can be identified by comparisons of allelic profiles from MLST data using the eBURST algorithm. Previous clustering methods such as dendrograms poorly represent recent evolutionary events and can be deceptive in their attempt to reconstruct deeper relationships. eBURST provides a more intuitive method representing the shorter-term relationships between sequence types (STs) based on a simple model for the emergence of 'clonal complexes'. Such complexes can occur when a founder sequence type rises to frequency within the population, either by a selective advantage or by random genetic drift to become a predominant clone. With this rise in frequency this ST eventually diversifies generating rarer close relatives which together form a clonal complex. Even in a population where recombination is frequent the strength of selection upon adaptive clones may result in the observable frequency of the clone superimposed on a background of genetic diversity which in time will be decayed by recombination (Figure 3).

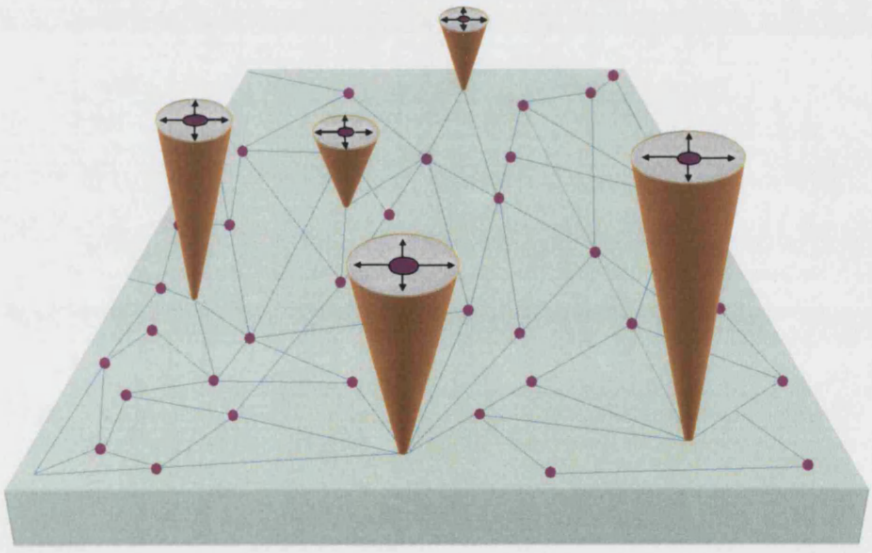


Figure 3. Illustration of the rise of clonal complexes upon a background of diversity.

These genotypes which have allelic profiles that differ from that of the founder ST at only one of the seven MLST loci, are called single locus variants (SLVs). These SLVs will eventually also diversify to produce variants at two of the seven loci (double locus variants; DLVs) (Feil *et al.*, 2004).

a

SEQUENCE TYPE	ALLELIC PROFIL E							ISOLATES	
ST45	10	14	8	6	10	3	2	13	FOUNDER
ST46	10	14	8	6	**14**	3	2	3	SLV
ST54	10	14	8	**13**	10	3	2	3	SLV
ST53	10	14	8	6	10	**16**	2	1	SLV
ST47	10	**11**	8	6	10	3	2	1	SLV
ST48	**17**	**11**	8	6	10	3	2	1	DLV

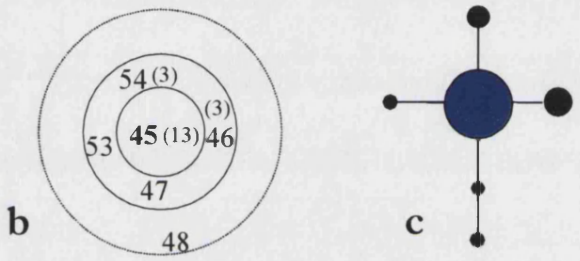


Figure 4. Representation of clonal complexes.

a) examples of closely related allelic profiles from *S. aureus* MLST data. b) representation of STs from 4a in a clonal complex (cc45) identified by BURST. The number in parentheses represents the number of isolates. The inner circle contains the founder ST. The second circle contains SLVs. The second circle contains DLVs. c) alternative representation of the clonal complex (cc45) by eBURST. Blue represents the founder ST. Black represents SLVs and DLVs. Circles are replaced by lines and the size of the circle indicates the number of isolates.

The reconstruction of such clones within a population facilitates the implementation of a method described by Feil *et al.*, 2003 which estimates the relative contributions of recombination and point mutation to this diversification of bacterial clones (Spratt *et al.*, 2001). This simple method measures two parameters within clonal complexes:

- the rate at which recombination changes an allele compared to point mutation.

- the rate at which recombination changes individual nucleotide sites compared to point mutation.

Within an SLV each variant allele that differs at multiple nucleotide sites from that of the founder (ancestral allele) is likely to have arisen by recombination since it is unlikely to have arisen by multiple independent mutation events whilst the remaining loci are unchanged. Point mutation will almost certainly result in the generation of a novel allele whereas recombination involving the complete allele will most likely be present in another isolate in the wild. However, an exception to this rule would be partial allelic recombination which would also generate a novel allele from donor and recipient sequences. With the characterisation of large numbers of strains by MLST it is likely that the majority of alleles that are present at significant frequency have been identified. However, the rate of recombination may be underestimated where there has been recombination between very similar sequences which results in only a single base change (Feil *et al.*, 2004).

The application of this method to *Staphylococcus aureus* suggests that both alleles (and individual sites) are 15 fold more likely to diversify by point mutation than recombination (Feil *et al.*, 2003). This contrasts with *Neisseria meningitidis* and *Streptococcus pneumoniae* where allelic changes by recombination are between 5 -10 fold more likely than those by point mutation. The ratio of recombination to mutation per individual sites was found to vary more significantly between *N. meningitidis* and *S. pneumoniae.* This presumably reflects differences in the sequence diversity of the two species whereby both donor and recipient alleles will typically be more variable in *N. meningitidis* (Spratt *et al.*, 2001).

The assessment of phylogenetic congruence between individual MLST loci can also be used as a measure of the impact of recombination. If a species is strictly clonal, trees generated from different loci should be congruent (in agreement) whereas trees from different loci will be incongruent if there is recombination. In the presence of recombination the loci of individuals within a species can resemble each other through the exchange of genes and parts of genes and not because they have a common ancestor. Congruent trees are therefore indicative of a relatively low rate of recombination to clonal divergence and are likely to illustrate the true relationships between strains, and subsequently the evolutionary history of the species. The level of congruence can be examined using a Maximum Likelihood (ML) method which scores individual loci trees against trees of random topology. If there is no phylogenetic congruence the individual loci trees will score no better than those of random topology. The ML trees for different MLST housekeeping loci for *N. menigitidis*, *S. pneumoniae* and *S. pyogenes* are no more similar to each other than they are to trees of random topology (Feil *et al.*, 2001). This loss of phylogenetic signal is consistent with the high estimates of recombination/mutation in *N. meningitidis* and *S. pneumoniae*. *Escherichia coli* was found to have a high level of congruence (Spratt *et al.*, 2001) agreeing with previous suggestions of a clonal structure in this species (Milkman & Bridges, 1990). *Staphylococcus aureus* has a more intermediate level of phylogenetic congruence with 55% significantly congruent pairwise comparisons (Feil *et al.*, 2003). These results suggest that, whereas the phylogenetic signal has been obliterated by recombination for *N. menigitidis* and *S. pneumoniae*, intraspecific phylogeny may still be feasible for *S. aureus*

1.4 Variation between different loci

As has been previously discussed, the frequencies of alleles observed within a species or population will be determined by the effects of random genetic drift and natural selection. However, the relative contributions of these two factors are largely unclear. Mutations have 3 possible selective outcomes: deleterious, neutral and advantageous. Whereas most biologists agree that deleterious mutations are the most frequent and will be purged by purifying selection, they disagree about the relative frequencies of neutral and advantageous mutations. The selectionist explanation emphasises the effects of natural

selection in producing variation by the fixation of advantageous mutations. However, this Darwinian explanation was challenged with the discovery of extensive genetic variation within and between species. This posed a problem for theories of natural selection which were thought to impose too high a selective cost on populations. This has been called the cost of natural selection which was proposed by J.B.S Haldane (Haldane, 1957). If there is extensive variation as a result of natural selection, then inferior alleles must be removed from the population by negative selection. The Neutral Theory of molecular evolution proposed by Kimura (1968) and King and Jukes (1969), overcame this problem by explaining variation as a result of mutation and random genetic drift fixing neutral mutations of little, if any, selective cost. Under this theory there are many alleles of equal value (Kimura, 1968; King & Jukes, 1969). However, the Neutral theory does not suggest that random drift explains all evolutionary change: natural selection is still needed to explain adaptation.

The distribution of variation can vary dramatically throughout a genome. Assuming a constant mutation rate and representative sample (avoiding sample bias), if there is little variation within a given gene, negative selection must have removed mutant alleles from the population. Most mutations therefore must have been deleterious to the function of this particular gene and its product. In this way some genes are purged of variation as it results in reduced or loss of function. In contrast, for some genes variation is maintained to observable frequency rather than selectively removed. In these cases it must either maintain or enhance function, or alternatively the loss of function has no impact on the reproductive potential of the individual. Since nonsynonymous substitutions are more likely to be deleterious the ratio of synonymous and nonsynonymous substitutions (d_S/d_N) can be used as a measure of the strength of purifying selection and subsequently functional constraint acting upon different genes. Such a measure has been used to compare the strength of selection between bacterial species (Jordan *et al.*, 2002).

The probability of a mutation being deleterious, neutral or advantageous is determined by the function of the gene it falls within (Figure 5). 16S rDNA codes for ribosomal RNA, part of the ribosome. This is a highly conserved gene as a result of the crucial role played by ribosomes in protein synthesis. This functional essentiality has massively constrained

the rate at which this gene evolves; mutations at this locus are likely to be lethal. Information pathway genes are typically maintained between bacterial genera and species. These genes are also involved in important cell functions such as DNA replication and transcription. Many information pathway gene products interact with the products of other genes and this is considered to confer a level of functional constraint (the complexity hypothesis) (Jain *et al.*, 1999). Housekeeping loci have been used for MLST schemes. These genes perform an essential metabolic function in the bacterial cell which is unlikely to be improved by substitution. Thus variation observed in these genes is most likely to be neutral. Deleterious nonsynonymous substitutions will eventually be removed by purifying selection. Variation at these loci is unlikely to be lethal. Genes which encode products exposed to an external environment may not be essential for cell survival but provide an adaptive function. In this case a relaxed functional constraint is preferential. Such proteins are more likely to be under a diversifying selection pressure since variation at such loci will not be lethal and may even confer an adaptive advantage. For surface exposed antigenic proteins the rapid and constant alteration of these proteins can promote evasion of the host's immune response.

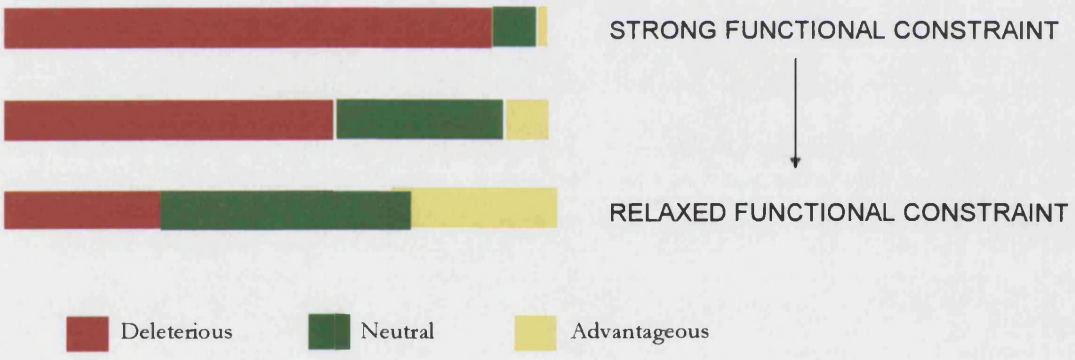


Figure 5. The fate of mutations relative to functional constraint.

Bacterial typing and gene choice

Recombination is commonly observed within genes where resulting recombinants have conferred a selective advantage. Recombination could be rare within a species as a whole but observed frequently with such a rise to high frequency of recombinants in a

population under strong selection. The intense selective pressures resulting from the continual exposure of *Neisseria gonorrhoeae* to penicillin, for the treatment of gonorrhoea, have resulted in the emergence of resistant strains with decreased affinity for the antibiotic. Comparative sequence analysis suggested that interspecific recombination had contributed to the acquisition of penicillin-resistant forms of PBP 2 (penicillin-binding protein) (Spratt, 1988). Therefore, selectively neutral loci will more accurately represent the stable genome 'backbone' in which recombinants will only rise to frequency by genetic drift. For this reason housekeeping genes are considered preferable for MLST schemes as they encode essential metabolic enzymes. Variation within these genes is likely to be selectively neutral or deleterious. However, the resolution afforded by any particular locus for a particular species will depend upon two variables: the number of informative sites within the specific locus and the extent of diversity within the species/ population to be examined as a whole.

Housekeeping loci provide little resolution within homogenous species such as *Bacillus anthracis* and *Mycobacterium tuberculosis* which correspond to a single lineage. These species may have emerged recently (Sreevatsan *et al.*, 1997) or undergone a recent population bottleneck (Parkhill *et al.*, 2003). In such cases more rapidly evolving loci are required to discriminate between strains and a scheme based on a variable number of tandem repeats (VNTR) has been developed for *B. anthracis*, known as MLVA (multiple loci VNTR analysis) (Hoffmaster *et al.*, 2002). At the other end of the scale, housekeeping MLST for highly diverse bacteria such as *Helicobacter pylori* resulted in the majority of isolates being assigned a unique sequence due to the high rate of recombination by natural transformation and a high mutation rate (Falush *et al.*, 2001; Suerbaum *et al.*, 1998). The uniformity within 16S rDNA loci is such that its use should be restricted to comparisons between genera (Figure 6). Providing no resolution at the species level, 16S rDNA has been used to infer much older phylogenetic relationships within the bacterial kingdom and the prokaryotic domain (Woese, 1987).

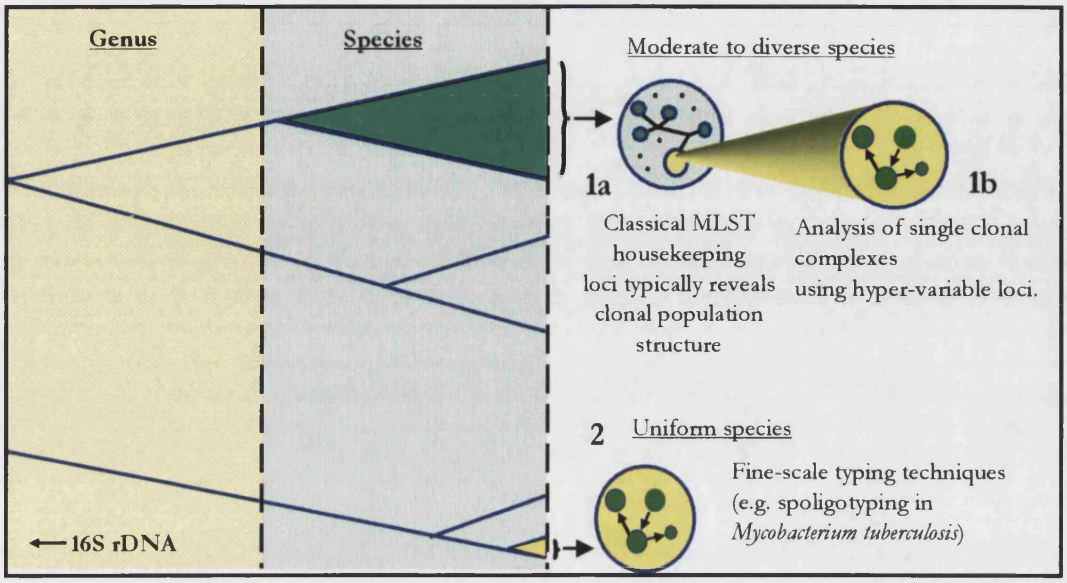


Figure 6. Gene choice for bacterial typing (Cooper & Feil, 2004).

1a) Species with moderate to high levels of diversity can be initially subdivided into clonal complexes on the basis of housekeeping gene variation.

1b) The variation within single clonal complexes (shown in red) can then be examined using more hyper-variable loci.

2) In highly uniform species a classical MLST scheme would reveal very limited diversity and such species can be considered the equivalent to a single clonal complex. Typing schemes in these species are thus restricted to the use of hyper-variable loci.

1.5 HGT in bacterial evolutionary history

Evidence for HGT across taxonomic boundaries is often cited based upon aberrant mutational biases, nucleotide and dinucleotide frequencies, codon usage bias and GC content (Karlin & Burge, 1995; Karlin *et al.*, 1998; Karlin & Mrazek, 2000; Lawrence & Ochman, 1997; Lawrence & Ochman, 1998; Moszer *et al.*, 1999). Evidence for lateral transfer can also be taken from incongruent phylogenies and gene content where genes present in isolated taxa are found to be absent in closely related species or genes with a high level of similarity are found in unrelated taxa.

Evidence for HGT has been found within the deepest branches of the tree of life. Cross domain transfer of plasmid DNA from *Agrobacterium tumefaciens* to plant cells in order to initiate tumor formation was one of the first observed HGT events (Chilton *et al.*, 1977; Stachel & Zambryski, 1986). Since then there have been many examples including homology within glucosyl hydrolases sequences of the Fungi (eukaryotes) and Bacteria (prokaryotes) (Garcia-Vallve *et al.*, 2000) and within DNA polymerase IV sequences of the Archeon *Methanosarcinia mazei* and bacteria (Deppenmeier *et al.*, 2002).

Accurate estimates for the impact of HGT are difficult. However, its impact is likely to have been underestimated since any different quantative estimates of various parameters: divergence time, amelioration rate (whereby mutational biases of the recipient genome result in the amelioration of foreign DNA which then becomes similar to the recipient), natural deviations and other factors affecting GC content. Thus ancient HGT events will become undetectable. Transfer of DNA between taxa of similar GC content may also be masked. The different taxonomic scale represented by many studies is also the source for much disagreement regarding HGT. In a group of closely related bacteria such as Enterobacteria, there may be congruence between single genes. Transfers occurring prior to the diversification of such a group can only be detected in larger phylogenetic reconstructions. These problems associated with accurate identification detection result in conflicting reports on the extent of HGT and its confounding effects in the reconstruction of organismal phylogeny (Doolittle, 1999; Kurland, 2000; Lawrence & Ochman, 2002). Although some authors still take a conservative view (Eisen, 2000; Logsdon & Faguy,

1999) others find little else to explain the apparent erosion of phylogenetic signal within conserved genes (Teichmann & Mitchison, 1999).

The impact of HGT on bacterial phylogeny

The existence of HGT greatly compromises one of the principle goals of evolutionary research: to reconstruct organismal phylogeny. Ribosomal RNAs (rRNA) have been considered the most informative phylogenetic markers and molecular clocks. They show a high degree of functional consistency, occurring in all organisms and there are different rates of change between different positions allowing the measurement of relationships at most phylogenetic levels (Woese, 1987). Thus, the discovery of lateral transfer of subunit ribosomal RNA (16S rRNA) (Ueda *et al.*, 1999; Yap *et al.*, 1999) was unsettling to the world of phylogenetics and lead to speculation regarding the existence of a shared 'core'. This 'core' would comprise a set of genes conserved at deep phylogenetic levels which have maintained phylogenetic signal through immunity to HGT. Naturally, with an increase in phylogenetic depth or taxonomic rank the number of shared genes decreases and notable cores of decreasing size have been observed (Harris *et al.*, 2003). Makarova *et al*, (1999) suggested that the 543 COGS (clusters of orthologs) shared between 4 euryarchaeal genomes (*Archaeoglobus fulgidus, Pyrococcus horikoshii, Methanobacterium thermoautotrophicum, Methanococcus jannaschii*) could be considered the 'evolutionary stable core' of the archaeal genomes. Genes involved in information pathways (transcription and translation) and those unique to archaea and eukaryotes were overrepresented within this group (Makarova *et al.*, 1999).

The functional essentiality of information pathway genes explains the conservation of such loci in distant relationships, although this is not evidence in itself for evolutionary stability. The 'complexity hypothesis' is often cited in support of this notion. This hypothesis suggests the formation of large complex systems with numerous interactions may confer a level of restriction and create a barrier to successful transfer of information pathway genes (Aris-Brosou, 2005; Jain *et al.*, 1999). In contrast to this idea Frederick Cohan suggested that extensive homologous recombination could actually restrict divergence between taxa in a positive feedback loop, whereby the diversifying effect of neutral mutation is balanced by the homogenising effect of recombination (Cohan, 1995). In this way, the limited

divergence of some conserved proteins could be a result of extensive recombination. Indeed, the frequency of recombination has been shown to decrease dramatically as sequences diverge (Harris-Warrick R. M., 1978; Majewski & Cohan, 1998; Majewski & Cohan, 1999; Zawadzki *et al.*, 1995). The overall divergence between a donor and recipient lowers the stability of the donor-recipient heteroduplex (an intermediate step in integration) and thereby lowers the probability of recombination (te Riele & Venema, 1982; te Riele & Venema, 1984). This can also be dictated by the machinery involved (Vulic *et al.*, 1997). *E. coli* is a largely clonal species with low levels of genetic exchange. Barriers to recombination between *E. coli* and *Salmonella typhimurium* are weakened in strains carrying mutations in genes encoding the methyl-directed mismatch repair system (MMR). Mismatches in the heteroduplex formed between donor and recipient DNA are identified by MMR and subsequently degraded (Rayssiguier *et al.*, 1989). Isolates which are MMR deficient have an enhanced rate of mutations and HGT. Despite its natural competency the *Haemophilus influenzae* genome bears little foreign DNA which is presumably a reflection of specific uptake sequences (Goodman & Scocca, 1988). Ecological barriers can also decrease the frequency for recombinational exchanges. *Mycoplasma genitalium* is also observed to have undergone little genetic transfer (Garcia-Vallve *et al.*, 2000). This species is an obligate intracellular parasite which may result in less opportunity for exposure to foreign DNA.

Natural selection or genetic drift will determine whether a recombinant becomes fixed and rises to observable frequency in the population. If natural selection is the arbiter of recombinational success then in functionally constrained loci where variation incurs a selective cost, recombination between diverged sequences is unlikely to be selectively tolerated. However, the lack of sequence divergence within such loci would conversely increase the likelihood of successful integration of donor DNA into a recipient genome. In the same way more divergent loci, where variation is promoted by diversifying selection, recombinant sequences are likely to be favoured. However, sequence divergence may be a limiting factor.

Several studies have attempted to establish which genes represent a conserved 'core'. Nesbø *et al*, use comparative phylogenetic analysis to examine the 'core' hypothesis in 512

COGS of the four euryarchaeons previously identified by Makarova *et al,*. (Makarova *et al.*, 1999; Nesbo *et al.*, 2001). A phylogeny of only four taxa has only 3 possible topologies. A conserved core of genes would be expected to consistently favour one topology over the other two. However, they found no evidence for the over-representation of a particular topology and fewer than half of the 521 COGS analysed carried a strong enough phylogenetic signal to be included in the analysis, of which only a third produced four taxon topologies. Further likelihood analysis of the distribution of information pathway genes between the three topologies revealed no significant difference. In this case these data do not support the 'core' hypothesis. Daubin *et al.*, 2002 implement a supertree approach to reconstruct a phylogenetic tree of 45 organisms representing three domains allowing the incorporation of non-ubiquitous genes and a larger data set to be considered. This approach still requires a conserved core set of genes upon which to build the supertree. Topologies for 310 gene trees were compared. Almost all information pathway genes were found to share, or at least partially share a common phylogenetic signal. Operational (housekeeping) genes were also represented within those with common signal although they displayed a broader range of topologies (Daubin *et al.*, 2002).

In favouring the supertree approach Daubin *et al.* (2002) highlight objections to the concatenation of gene sequences due to the weighting of genes within a dataset whereby many sites may be contributed by just a few genes/proteins. This may have a particularly significant effect if these are the genes which are found to have undergone transfer. However, with a sufficiently large gene set this problem may be overcome. One hundred and six yeast genes in seven *Saccharomyces* sp. and *Candida albicans* were concatenated in an attempt to overcome the incongruencies of several branches within single loci phylogenies. Concatenated sequences yielded a tree with 100% bootstrap support for each branch irrelevant of the method of phylogenetic construction (Maximum likelihood on nucleotides/Maximum parsimony on nucleotides/Maximum parsimony on amino acids). The maximum support attained for a single topology in this study is strongly supports the power of this method in overcoming the incongruence present in single gene analyses.

There appears to be little consensus regarding the existence of a 'core' gene set and the transferability of conserved information pathway genes. The study of Nesbø *et al* contrasts

with other studies where the transfer of information pathway genes appears to be preferentially curtailed (Jain *et al.*, 1999). However, this may be explained by the problems previously identified in estimating and accurately identifying lateral transfer: these two studies cover different taxonomic boundaries. Whereas Nesbø *et al* examine four euryarchaeal genomes, Jain *et al.* examined orthologs of six more distantly related genomes (*Escherichia*, *Methanococcus*, *Bacillus*, *Aquifex*, *Archaeoglobus*, *Synechocystis*). One explanation for this discrepancy would be that transfer between distant taxa (Bacteria and Archaea) is indeed severely constrained (by failure of their products to integrate with resident components of complex information processing systems). For close (albeit still quite different) taxa such as the Euryarchaea, rates of transfer of Information pathway and Housekeeping (operational) genes might be more nearly equal. In other words information pathway genes might be easily transferred within, but not between domains.

Intraspecific phylogenies have previously been described for many bacterial species utilising different types of genes where the locus choice has often been determined by limited sequence data availability for primer design or clinical interest. There is little concurrence among the genes examined across species. Santos and Ochman have developed primer sets in an attempt to reconcile these differences and provide the scientific community with conserved gene sets with which to both identify and characterise bacterial species (Santos & Ochman, 2004). Single loci phylogeny may be misleading for inferring relationships between strains, and conflicts between single loci phylogenies for bacterial species have been reported. Incongruence between housekeeping loci, considered to be representative of the neutral genome is observed in recombining species illustrating the conflicts in phylogeny provided by different loci. However, accurate phylogeny, particularly for clinically relevant bacterial species, plays a central role to epidemiological studies and a greater understanding of the dissemination of clones. This is also important for the understanding of disease potential. The generation of a well supported tree for the clonal species *E. coli* revealed the parallel evolution of virulence determinants (Reid *et al.*, 2000).

1.6 Biology of *Staphylococcus aureus*

Staphylococcus aureus is a coccoid, cluster-forming Gram-positive bacterium which forms part of the natural microflora of humans and animals. Although found on skin and mucous membranes, *S. aureus* predominantly resides within the anterior nares (Peacock *et al.*, 2001). It is estimated to be asymptomatically carried in a healthy adult population by ~20% of individuals persistently and by ~60% intermittently (Peacock *et al.*, 2001; von Eiff *et al.*, 2001). However, it is an opportunistic pathogen but of no specific disease. It can cause suppurative infections such as furuncles and styes, exfoiliative dermatitis, mastitis and deep tissue abscesses. *S. aureus* is also responsible for systemic infections such as bacteremia, osteomyelitis and pneumonia. It can also cause toxic shock syndrome and food poisoning. It was first described in 1880s by Sir Alexander Ogston at the Aberdeen Royal Infirmary who found it was the most common cause of infected surgical wounds (Ogston, 1882).

The development of drug resistance in *S. aureus*

In a world before antibiotics bacterial infections such as tuberculosis and pneumonia were infamous killers and 80% of *S. aureus* bacteraemias were fatal (Skinner, 1941). The discovery of penicillin in the late 1920s and its development for antibiotic treatment by Ernst Chain and Howard Florey was to save thousands of lives. Penicillin prevents the cross-linking of peptidoglycan in the cell wall so that malformed cells undergo osmotic rupture. The majority of soldier deaths in World War I were not on the battlefield but due to bacterial infections. Penicillin was introduced in 1944 during World War II and many soldiers' lives were saved. Socio-economic developments including sanitation also contributed to the increase in life expectancy in the 20th century as fatal bacterial diseases became curable. However, as some were ready to close the book on bacterial infections, even Alexander Fleming himself warned that with improper use "microbes could be educated to resist penicillin..." Nevertheless, the use of antibiotics was heavy and often inappropriate. Despite the warning of Alexander Fleming, only 4 years after its introduction, half of all hospital *S. aureus* isolates were resistant to penicillin through inactivation by penicillinase (β-lactamase) (Barber & Rozwadowska-Dowzenko, 1948). Other natural antibiotics were developed including chloramphenicol, erythromycin,

streptomycin and tetracycline. Resistance emerged again rapidly. Hope came in the form of penicillinase-stable β-lactams such as cephalosporin and subsequently the semi-synthetic penicillin, methicillin. Alas, the first methicillin-resistant *S. aureus* (MRSA) were seen in 1961, the very first year of methicillin usage (Jevons, 1961). Methicillin resistance is conferred by the *mecA* gene carried on a large section of chromosomally inserted DNA termed staphylococcal cassette chromosome *mec* (SCC*mec*). *MecA* encodes a novel penicillin-binding protein (pbp2' or pbp2a) which has a reduced affinity for β-lactam antibiotics (Brown & Reynolds, 1980; Hartman & Tomasz, 1984). The presence of this protein allows the crosslinking of peptidoglycan even when the normal pbps (1, 2 and 3) have been inactivated. The use of alternative antibiotics such as gentamicin curbed the rise in MRSA spread for a time but the spread of gentamicin-resistant MRSA became a problem in the early 1980s with the beginning of 'epidemic' strains (Anon, 1995). The most prevalent MRSA strains in the UK are EMRSA-15 (ST22) and EMRSA16 (ST36). The remaining class of antibiotics where susceptibility is still high is the glycopeptides (teicoplanin and vancomycin). The first *S. aureus* with reduced susceptibility to vancomycin (>8mg/l) was isolated in Japan in 1997 (Hiramatsu *et al.*, 1997; Hiramatsu *et al.*, 2002). More recently there have been reports of further clinical strains which carry vancomycin resistance genes (*van* genes) which confer vancomycin resistance in the enterococci (CDC, 2002a; CDC, 2002b; CDC, 2004). Resistance is conferred by the thickening of the cell wall which effectively prevents vancomycin molecules from reaching the cytoplasmic membrane by trapping them within the pre-existing cell wall (affinity trapping) (Cui *et al.*, 2000).

The spread of MRSA, and increases in nosocomial infection have made aspects of bacterial evolution a topic of both clinical and political relevance. It is therefore urgent to improve our understanding of the mechanisms responsible for resistance dissemination and to continue the development of alternative therapies and containment strategies for this adaptive pathogen.

Why S. *aureus* is such a successful opportunistic pathogen

The ability to resist antibiotic treatment is not in itself a disease causing factor, it simply makes therapy problematic. The range of disease that *S. aureus* can cause is due to a

plethora of virulence-associated factors summarised in figure 7, some of which will be described here. *S. aureus* adherence to host cell proteins such as collagen, keratin and fibronectin which form the extracellular matrix of epithelial and endothelial surfaces, is facilitated by the action of an array of surface proteins. Most *S. aureus* strains express both fibronectin and fibrinogen binding proteins which promote attachment to blood clots and traumatised tissues. Interactions with collagen may also be important in promoting bacterial attachment to damaged tissue where the underlying layers have been exposed. Once bacterial attachment has occurred other staphylococcal components enable the avoidance of host immune responses. The expression of a capsular polysaccharide (or microcapsule) can impede phagocytosis. The surface protein, protein A can bind host IgG thus disrupting opsonisation and phagocytosis. *S. aureus* is coagulase-positive protecting bacterial cells from phagocytic and immune defenses by causing localised clotting. Subsequent toxigenesis is mediated by an array of staphylococcal toxins. The staphylococcal enterotoxins (SEs) can cause staphylococcal food poisoning and can also cause toxic shock when expressed systematically. The toxic shock syndrome toxin (TSST-1, previously *sej*) is an exotoxin associated with toxic shock syndrome (Bergdoll *et al.*, 1981). These toxins have superantigenic activity stimulating the production of large T-cell populations and subsequently large amounts of cytokines which are responsible for the symptoms of toxic shock. Further exfoliatins can be released and are associated with scaled-skin syndrome where there is widespread blistering and loss of the epidermis. A pore forming cytotoxin, Panton-Valentine leukocidin (PVL), targets mononuclear, polymorphonuclear and epithelial cells inducing severe inflammatory lesions and skin necrosis (Prevost *et al.*, 2001; Ward & Turner, 1980). Cases of community-acquired MRSA (CA-MRSA) due to PVL-positive *S. aureus* have been reported in Europe and the UK (Klein *et al.*, 2003). There has been much interest most recently in PVL due to the incidence of severe disease in children and young adults and the association between PVL and community acquired pneumonia (Lina *et al.*, 1999). *S. aureus* also excretes other extracellular enzymes such as proteases, a lipase, a deoxyribonuclease and a fatty acid modifying enzyme which are involved in the destruction of host tissues and in the provision of nutrients to infecting bacterial cells.

The hospital environment offers many opportunities to such a versatile pathogen and hospital-associated infections are common. In the absence of the highest hygiene standards, there is the potential for transmission between immunologically weakened hosts. The opportunity to breach physical barriers, via intravenous lines, ventilation tubes and open wounds, and cause disease is frequent. Implanted surfaces such as catheters and replacement hips but also heart valves are vulnerable to colonisation post surgery and thus treatment of this type of infection can be highly problematic.

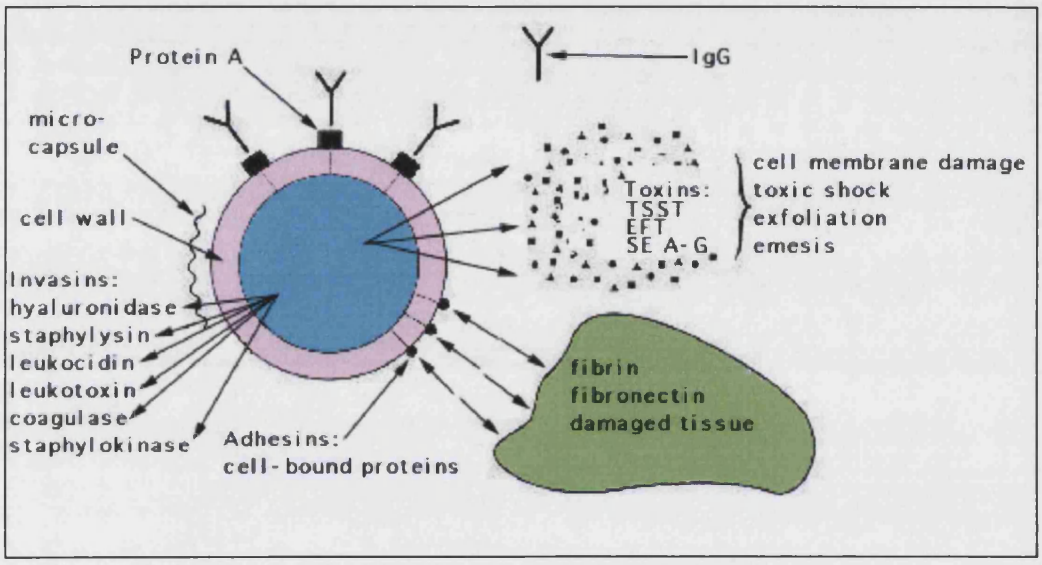


Figure 7. Illustration of the array of virulence-associated factors of *Staphylococcus aureus* (taken from http:textbookofbacteriology.net/staph.html)

Population structure of *Staphylococcus aureus* from 'core' variation

The short term evolutionary history can be analysed using MLST data and is represented below using eBURST. The *Staphylococcus aureus* MLST database now contains approximately 1400 isolates (Simon O' Hanlon, July 2005, personal communication).

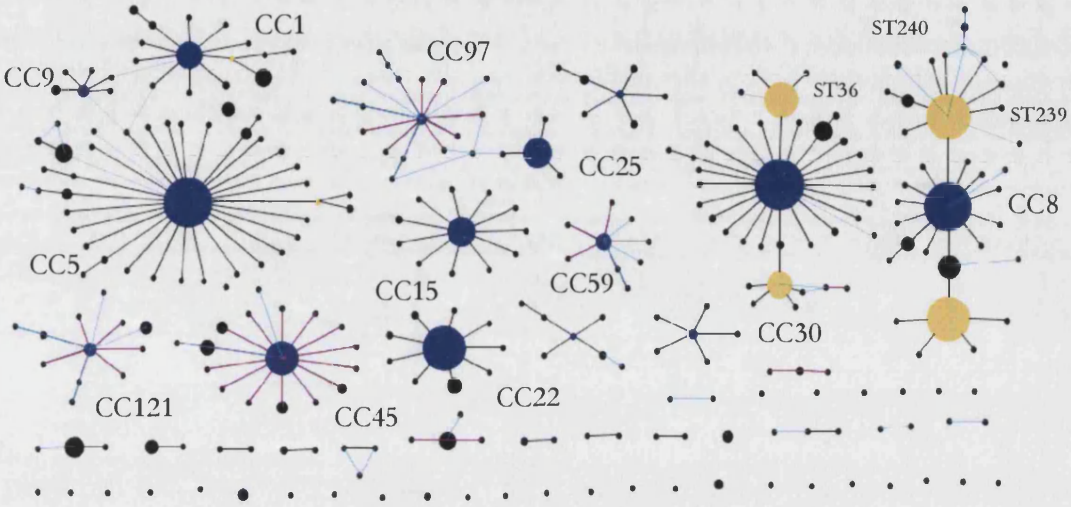


Figure 8. A population snapshot of 1243 (May 2005) isolates in the *S. aureus* MLST database using eBURST.

The natural population structure of *S. aureus* has been analysed from the housekeeping sequences of MLST data. This was conducted on a large collection of 334 isolates from Oxfordshire, UK. These are recovered from asymptomatic carriage in the anterior nares, community-acquired disease and hospital acquired disease. This collection also includes both methicillin-sensitive and methicillin-resistant isolates represent most clonal complexes shown in Figure 8. Most of these clonal complexes fit a simple model of radial diversification from a clonal founder. This study by Feil *et al.* in 2003 confirmed previous studies by MLEE (Musser *et al.*, 1990) that *S. aureus* is a highly clonal species (Feil *et al.*, 2003). It was also observed that disease isolates were equally distributed among clonal complexes indicating that there is no link between clonal background and the propensity to cause disease. However, these findings do not equate to an equal virulence potential in *S. aureus*. These findings demonstrate that carried and disease causing isolates are indistinguishable based on the 'core' variation afforded by ubiquitous housekeeping loci.

1.7 The staphylococcal 'accessory' genome

There is another gene pool within bacterial genomes which is comprised of non ubiquitous 'accessory' genes. These may not be present in all strains either between or even within bacterial species. They are often adaptive loci such as drug resistance determinants or factors involved in host colonisation or invasive potential. Genome scale differences can be observed by comparative genomics between mrsa252, a hospital-acquired representative of the EMRSA-16 (ST36) and an invasive community acquired strain mssa476 (ST1). Although they have a well conserved core region they differ markedly in their accessory genetic elements. Both have the SCC element integrated within their genomes, however in mrsa252 it carries an array of resistant determinants including the methicillin-resistance determinant, *mec*, which are absent on the SCC element found in mssa476. These STs are found within different groups of the population however even between two strains of the same ST, mssa476 and MW2 there are differences in gene content. These strains are indistinguishable by genotype but differ by three large scale differences. Mssa476 lacks the SCCmec type IV cassette, the SaP13 pathogenicity island and the bacteriophage ΦSa2 (MW2) but contains a SCC – like element, SCC_{476} and bacteriophage ΦSa4 (476) (Holden *et al.*, 2004). Genome sequences also revealed *S. aureus* specific loci that were not present within the genome of *S. epidermidis* genome (ATCC12228). The absence of virulence associated factors such as leukocidins, superantigen genes and the majority of proteases and several lipoproteins correlate with the less aggressive pathogenesis of *S. epidermidis*. It had originally been speculated that methicillin-resistance may have arisen in a single descendent of all current MRSAs (Kreiswirth *et al.*, 1993). However, further studies have shown that some MRSA are very different (Fitzgerald *et al.*, 2001; Musser & Kapur, 1992; Sreevatsan *et al.*, 1997). MLST data and the resolution of clonal complexes was used to examine the clonal distributions and relationships between an international collection of MRSA isolates (n=359) compared with their sensitive counterparts (n=554) (Enright *et al.*, 2002). The short-term evolutionary relationships between isolates were established from MLST data by eBURST in Figure 8. Used in concert with SCC*mec* typing technique, the authors were able to present the putative evolutionary pathways for the emergence of the major EMRSA clones. The authors also clarified the repeated emergence of MRSA clones from successful

epidemic MSSA strains representing 5 clonal complexes. The presence of different SCC*mec* types within the same clonal complex supports the notion that horizontal transfer has been significant in the dissemination of methicillin-resistance in the past 40 years (Enright *et al.*, 2002). We can observe the distribution of MRSA isolates in the context of all identified *S. aureus* clonal complexes in Figure 9.

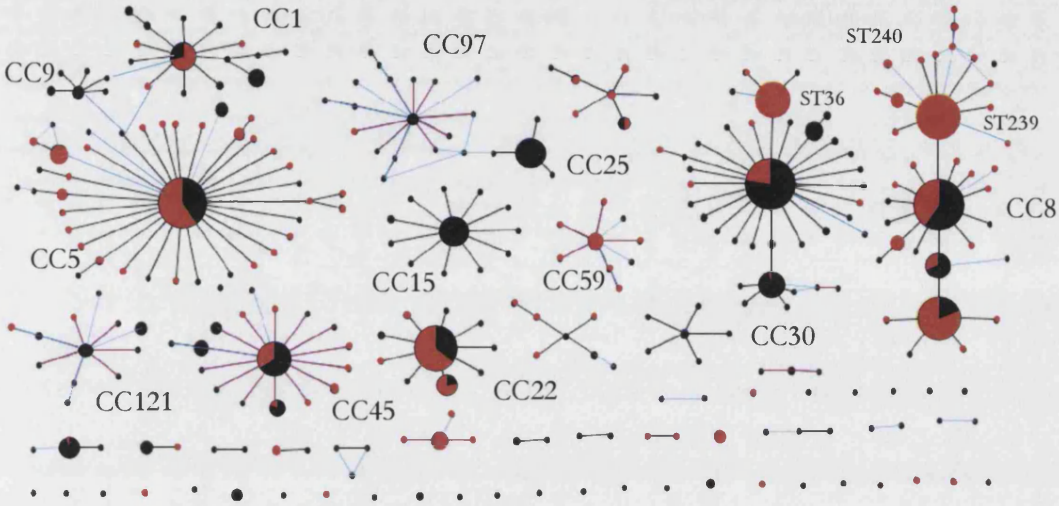


Figure 9. The distribution of methicillin-resistance within 1243 isolates of the *S. aureus* MLST database (May 2005). Resistance: Red, Susceptibility: Black.

The distribution of methicillin-resistance is clearly observable throughout the population and the presence of resistance is variable between clonal complexes, within clonal complexes and even within strains of the same ST. With the unique opportunity to observe bacterial evolution and adaptation in matter of decades, the short-term relationships within clonal complexes are an appropriate scale in which to observe the rapid acquisitions of methicillin resistance. However, selective pressures for rapid change of movement of adaptive loci are rarely observed in this way and a longer term population framework will be essential for examining the transmission of virulence-associated 'accessory' loci in this species. A study by Peacock *et al.* examined the distribution of 33 putative virulence determinants within the same Oxford collection. Their data clearly demonstrates the strain-specific distribution of many of the virulence-associated factors

within the *S. aureus* population. They also find a cumulative relationship with disease isolates whereby the more virulence-associated factors expressed, the more likely to isolate is to have been recovered from disease (Peacock *et al.*, 2002). Together these studies reveal that the propensity of *S. aureus* to cause disease is independent of 'core' backbone variation and linked to a profile of strain specific virulence factors.

MSCRAMMS and *Staphylococcus aureus* pathogenesis

Successful attachment to host surfaces is an essential prerequisite for colonisation and disease. Indeed one of the major risk factors for disease is nasal carriage. The rate of infection in carriers is much greater than in non-carriers (Weinstein, 1959) and individuals are typically infected by their own carried isolate (Luzar *et al.*, 1990; Nguyen *et al.*, 1999; Yu *et al.*, 1986). The temporary elimination of nasal carriage has also been shown to reduce the rate of nosocomial infection in dialysis and surgical settings (Boelaert *et al.*, 1993; Kluytmans, 1998; Yu *et al.*, 1986) The cell wall of *S. aureus* expresses a plethora of anchored proteins which enable cells to adhere to components of the extracellular matrix in order to initiate colonisation. Adherence is mediated by protein adhesins of the MSCRAMM (microbial surface components recognizing adhesive matrix molecules) family. These are characterised by the presence of a C-terminal LPXTG motif.

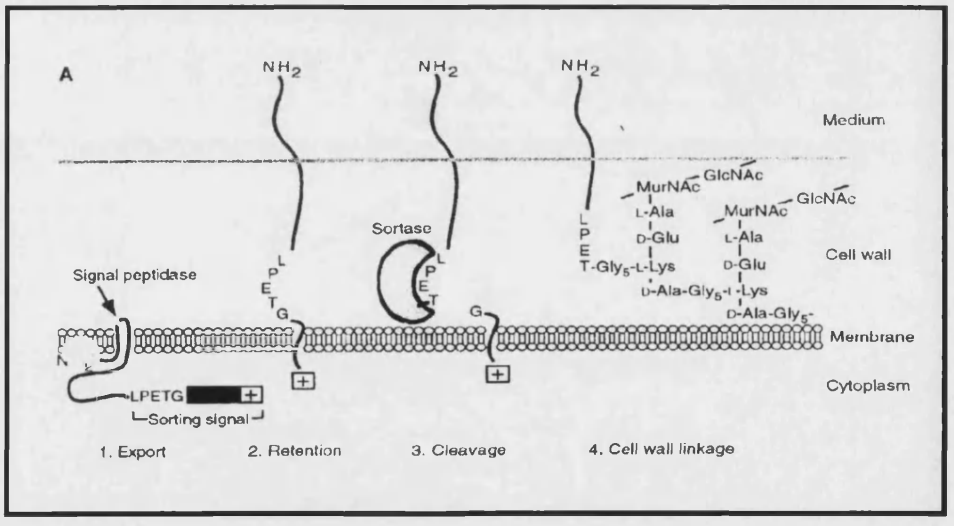


Figure 10. Anchoring of surface proteins in gram-positive bacteria.

The ligand binding function of MSCRAMMS is found in the amino-terminal domains whereas the carboxyl terminal typically contains a wall spanning region. The anchoring

mechanism requires an LPXTG motif followed by a hydrophobic domain and positively charged residues. Surface proteins are first synthesised in the bacterial cytoplasm and then transported (1. export) across the cytoplasmic membrane. The N terminal signal peptide of the cytoplasmic surface protein is cleaved (3. cleavage) generating the extracellular P2 species, which is the substrate for the cell wall anchoring reaction. Sortase, a membrane anchored transpeptidase, cleaves P2 between the theonine (T) and the glycine (G) of the LPXTG motif and catalyses the formation of an amide bond (4. cell wall linkage) between the carboxyl group of threonine and the amino group of cell wall cross bridges

Gene	prevalence %	Ligand	Other function	References*
spa	92	IgG	activation of platelet aggregation	1
cna	41	collagen	-	2
fnBPA	92	fibronectin, elastin, fibrinogen	-	3
fnBPB	-	fibronectin, elastin	-	4
clfB	100	fibrinogen, cytokeratin	activation of platelet aggregation	5
clfA	99	fibrinogen	activation of platelet aggregation	6
sdrC	100	putative adhesin	-	7
sdrD	41	putative adhesin	-	8
sdrE	48	putative adhesin	activation of platelet aggregation	9
bbp	41	bone sialoprotein		10
ebps	65	elastin	-	11
map/eap	94	fibronectin, fibrinogen, prothrombin	-	12

Table 1. *S. aureus* MSCRAMMS.

* 1.Uhlen *et al.*, 1984; O'Brien *et al.*, 2002 2. Patti *et al.*, 1992 3. Greene *et al.*, 1995; Jonsson *et al.*, 1991; Wann *et al.*, 2000; Roche *et al.*, 2004 4. Greene *et al.*, 1995; Jonsson *et al.*, 1991; Roche *et al.*, 2004 5. Ni Eidhin *et al.*, 1998; O'Brien *et al.*, 2002 6. McDevitt *et al.*, 1994; O'Brien *et al.*, 2002 7 - 9 Josefsson *et al.*, 1998 10. Tung *et al.*, 2000 11. Downer *et al.*, 2002 12. Palma *et al*, 1999.

The presence MSCRAMMS in the large Oxford collection of isolates representing disease and carriage is shown as a percentage presence when tested using a PCR assay (Table 1) (Peacock *et al.*, 2002). This demonstrates the diversity of *S. aureus* ligands and that adherence to any single component can be mediated by several surface proteins (functional redundancy) and a single surface protein often has multiple functions. Some MSCRAMMs have been well chacterised but analysis of multiple *S. aureus* genomes reveals the presence of further 10 uncharacterised proteins with LPXTG motifs (Roche *et al.*, 2003a). The ability of *S. aureus* to adhere to such a diverse array of its host's extracellular

matrix components no doubt contributes to its invasive potential and non-specific disease. However an interpretation of this data in the context of the population framework is required in order further elucidate the mechanisms which explain the distribution of these 'accessory' virulence associated factors in the population. The same study also showed significant associations between 3 of the above surface proteins, fnbA, cna and sdrE with isolates from disease. Of particular interest is the association of sdrE with disease and not bbp. Bbp was identified as a bone-sialoprotein binding protein (Tung *et al.*, 2000) and a close relative of sdrE. However, inspection of genome sequences reveals that *bbp* is found instead of *sdrE* at the same locus in mrsa252.

The SDR family

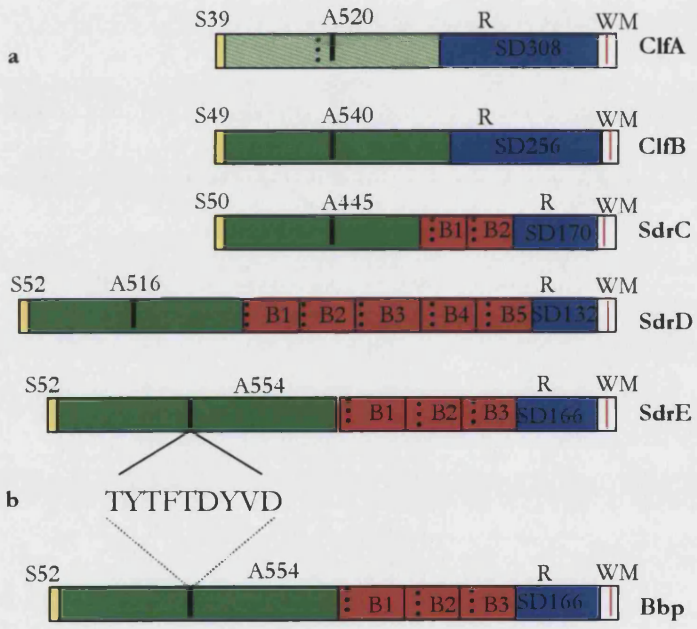


Figure 11. Structural organisation of the SDR proteins.

Although MSCRAMMS have similar structures, 5 of these genes are sub-characterised by regions of serine aspartate (SD) repeats at their C terminal end and are named the SDR family. Well characterised members of this family are the fibrinogen-binding proteins clfA and clfB (clumping factors). However, the less well characterised sdrC, sdrD and sdrE

form a subfamily within the SDR family since they have 2-5 additional B motifs between the A and R regions (Josefsson *et al.*, 1998) and are illustrated in Figure 11. In region A the thick black line represents the conserved TYTFTDYVD motif. In the region A of clfA and in regions B the thin dashed line represents an EF hand loop. S, signal sequence: A, putative ligand binding A region; B, B repeat; R, SD repeats; W, short wall spanning region; M, membrane spanning segment. The LPXTG motif occurs between domains W and M. Structurally the bbp protein is identical to the sdrE protein with the same number of B repeats and the same size A region.

Although *sdrE* and *bbp* have previously been considered separate genes, it is now clear that this is not the case. Based on the evidence for structural and sequence homology between *sdrE* and *bbp*, and the fact that they are present at the same locus, these shall be considered as '*sdrE*' and '*bbp*' type alleles for the remainder of this thesis.

The sdrC, sdrD and sdrE proteins are less well described than, the clumping factor proteins (clfA and clfB), and ligands are yet to be identified. The only described role for sdrE is the activation of platelet aggregation and the 'bbp' type allele has been reported to bind bone-sialoprotein (O'Brien *et al.*, 2002a; Tung *et al.*, 2000). However, in order to interpret variation at the sdrE locus and gain insight into putative function we can relate to the clumping factor proteins. ClfA and clfB are well characterized members of this family and are considered model proteins for all structurally related proteins: FnbpA, FnbpB, SdrC, SdrD, SdrE and cna. The A regions (putative binding regions) of these proteins are similar in size and are composed of significantly similar amino acid sequences. ClfA was the first of the fibrinogen binding proteins to be identified, isolated (Espersen & Clemmensen, 1982; Espersen *et al.*, 1985) and characterised at a molecular level (McDevitt *et al.*, 1994; McDevitt *et al.*, 1995; McDevitt *et al.*, 1997). Studies on ClfA localized the binding region, to within the C terminal end of the A region and has since been more specifically localised to potential residues which may be involved in ligand recognition. The fibrinogen binding activity of ClfA has been localised to residues 221-559 within the A region of this protein. In addition, the C terminal part of the A region (residues 484 - 550) has also been implicated as being important for Fg binding (McDevitt *et al.*, 1997). The use of site directed mutagenesis identified two adjacent residues Glu^{526} and Val^{527} as

being important for ClfA activity. *S. aureus* expressing a substitution at either site exhibited reduced ability to bind soluble fibrinogen and to adhere to immobilised fibrinogen. Further more those expressing both were almost completely defective in Fg binding (Hartford *et al.*, 2001b) The repeat (R) region is required for functional expression of the fibrinogen binding domain and its role in this protein is to span the entire wall and to display the biologically active A domain in a form that can participate fully in fibrinogen binding (Hartford *et al.*, 1997; McDevitt & Foster, 1995). The second of the *S. aureus* fibrinogen-binding proteins (clfB) was identified by hybridisation to a probe for the dipeptide repeat (Ni Eidhin *et al.*, 1998). However, whereas clfA binds exclusively to the γ-chain of fibrinogen, clfB binds both α- and β-chains of fibrinogen (Ni Eidhin *et al.*, 1998; O'Connell *et al.*, 1998) Localised binding activity in ClfB has also been investigated as a model for structurally related surface proteins. The presence of a cleavable SLAVA motifs indicated that the A region may be composed of subdomains.

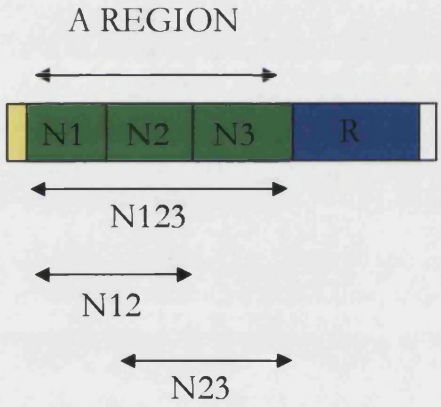


Figure 12. Putative subdomains of clfB.

The truncated recombinant N23 binds fibrinogen with the highest affinity whereas N12 and N123 bind with lower affinity. Individual subdomains of N1, N2 and N3 did not bind to immobilised fibrinogen. These results suggest that the protein segment required for a fibrinogen binding site is not contained in an isolated subdomain but that two subdomains (N12 or N23) are needed to form an active protein. Antibodies raised against N2 were the most effective at inhibiting the binding of *clfB* mediated adherence to fibrinogen

suggesting this domain plays an important role. Likewise, N23 effectively inhibited adherence to fibrinogen by *clfB* expressing *S. aureus* cells. The full length A region N123 also caused some reduction in bacterial attachment whereas N12 was essentially inactive. Thus, a recombinant N23 appears to bind fibrinogen with a higher affinity compared with that observed for N123. Partial removal of the N1 subdomain lead to an activation of fibrinogen binding activity, whereas complete removal of the N1 subdomain results in loss of fibrinogen-binding activity, since cleaved N23 does not bind fibrinogen (Perkins *et al.*, 2001). The metalloprotease aureolysin is responsible for this cleavage of the clfB A region. The biological implications of an MSCRAMM composed of subdomains which may be proteolytically processed are unclear. The functions of proteases are poorly understood, but there is evidence that they can cleave both bacterial and host proteins The loss of an N-terminal subdomain from the surface-associated protein may release biologically active peptide fragments. Alternatively, cleavage of clfB may promote the detachment of bacterial cells from colonised sites and facilitate the spread of infection within the host. However, regulated processing of N1 could greatly affect the fibrinogen binding activity of the remaining N23 domains of *clfB*. Truncated A regions have been tested for binding to immobilised human epidermal keratin. Both N123 and N23 bound to keratin in a dose dependent manner, whereas N12 did not bind. The notion that the keratin binding site may also be in the N23 subdomain is supported by almost complete inhibition of adherence to immobilised keratin by clfB expressing *S. aureus* cells (O'Brien *et al.*, 2002b).

A further role has been reported for the clumping factors clfA and clfB. Along with SDR family member sdrE and protein A, clfA and clfB were shown to activate the aggregation of platelets. *Lactococcus lactis* is gram-positive bacterium which can be used as a surrogate host for the assessment of candidate genes in several adherence assays (Hartford *et al.*, 2001a; O'Brien *et al.*, 2002b; Roche *et al.*, 2003b). Candidate genes were expressed in the non aggregating *Lactococcus lactis* and tested for the ability to stimulate platelet activation in platelet rich plasma (O'Brien *et al.*, 2002a). The residues involved in such a role are unknown. There are 3 platelet receptors via which protein A can activate the aggregation of platelets either the Fc receptor via and IgG bridge, with receptor GPIb (Hartleib *et al.*, 2000), via von Willebrand factor or by interacting with the platelet gC1qR/p33 receptor directly. ClfA and clfB may also interact directly with the platelet receptor P118 or with

receptor GPIIb/IIIa via a fibrinogen bridge (O'Brien *et al.*, 2002a). The sdrE interaction did not occur with gel filtered platelets but required the presence of plasma. Therefore this activation occurs via at least one unknown plasma protein with an unknown platelet receptor. The ability of the sdrE locus variant bbp to activate the aggregation of platelets has not been reported, nor has the ability of the 'sdrE' form of the locus to bind bone sialoprotein. Thus the functional implications of allelic variants at this locus are unclear.

Of further interest is the presence of the clumping factor B gene (*clfB*) in close proximity of approximately 1.5 kb to the *S. aureus* MLST housekeeping gene *arcC* (figure 14). Data from *S. aureus* MLST suggests a low level of recombination within housekeeping genes. Despite this there are some incongruencies between MLST loci. *arcC* is the least congruent of all the housekeeping genes for MLST and poorly supports a putative population division identified in remaining *S. aureus* MLST genes (Feil *et al.*, 2003).

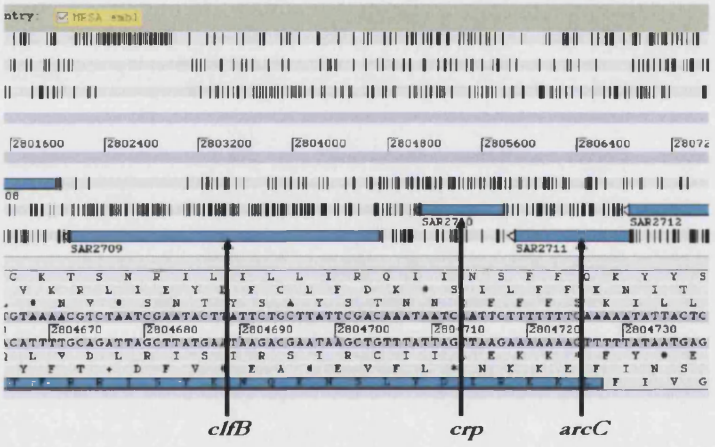


Figure 13. Genomic arrangement of *arcC* relative to *clfB*.

It is possible that the incongruence seen for the *arcC* housekeeping gene is a result of proximity to an SDR gene, *clfB* As has previously been described; the *clfB* protein interacts directly with its host. This may put this locus under diversifying selection in attempt to overcome host recognition. Alternatively, the 'accessory' nature of some of the functions of this protein may result in a relaxed functional constraint which case recombinant sequences are subject to a much reduced purifying selection. The 'hitchhiking' effect is defined as the changes in frequency or sequence evolution of a gene owing to its close

proximity ('linkage') to another gene under strong selection. The evolution at this locus is effectively a side effect of the evolution of another gene which is subject to stronger selective forces. Hitchhiking has been documented in several species. A high level of divergence within alleles of the housekeeping gene *ddl* of *Streptococcus pneumoniae* is attributable to recombinational events at the *pbp*2 locus driven by penicillin usage (Enright & Spratt, 1999). The ability to avoid host recognition confers a strong selective advantage and the hitchhiking effect is seen in the flanking regions of the O-antigen of *Salmonella* and *Escherichia coli* (Milkman *et al.*, 2003; Reeves, 1993; Reeves, 1995). It was found, that in a single transformation step, strains of *Streptococcus pneumoniae* could switch both capsular type and resistance profile with a great potential selective advantage (Trzcinski *et al.*, 2004).

1.8 Aims of this thesis

A large population scale reassessment of the impact of homologous recombination in housekeeping genes has already been undertaken using MLST sequence data for *Staphylococcus aureus*. In this thesis I present a comparative study using the characterisation of 'core' genome variation to assess the modes of nucleotide evolution in different functional categories of genes, including housekeeping genes, within the *S. aureus* genome.

The clinical relevance of this microorganism and the availability of a large collections of isolates from all epidemiological settings make this an ideal model microorganism for the appraisal of the link between functionality, diversity, recombination and phylogenetic consistency at the intraspecific level. Concatenated MLST data has been used in phylogenetic reconstruction in an attempt to resolve deeper evolutionary relationships between lineages of *S. aureus*. Although clonal complexes are readily identifiable on the tree in figure 15 and some terminal branches are resolved, the relationships between clonal complexes are less clear. Inspection of individual MLST loci trees suggest a conserved node dividing the population into two groups (Feil *et al.*, 2003), and although this may be real it is poorly supported. The internal branches of this tree are also poorly resolved and not well supported. The uppermost region of the tree is multifurcating. The utility of such a phylogeny is limited. As it stands, the phylogeny has little power as tool for the identification of branches of importance in the evolutionary history of the species, tracing

acquisition events of 'accessory' elements such as pathogenicity islands, drug resistance determinants and virulence-associated loci.

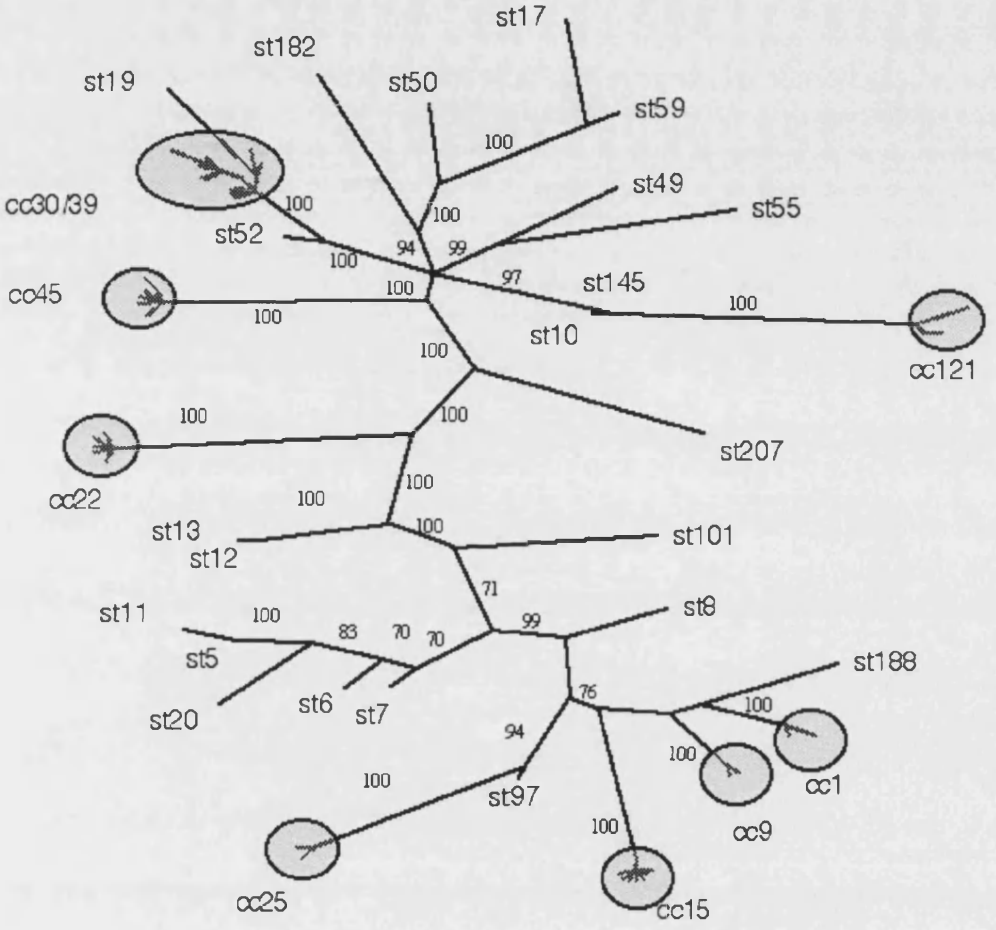


Figure 15 . Unrooted Bayesian tree of concatenated sequences of the 75 different STs of *S. aureus* as reported in Feil *et al.*, 2003.

A large sequence dataset representing 38 unlinked ubiquitous *S. aureus* genes in a diverse sample of the natural population will be used in the reconstruction of a robust phylogeny for this species. The individual MLST loci show an intermediate degree of phylogenetic congruence with each other. This is suggestive that there is some maintenance of phylogenentic signal in this species and with the inclusion of further data a robust and well supported phylogeny is feasible. This is supported by the evidence for relatively low rate of recombination in this species (Feil *et al.*, 2003).

Such a phylogeny will provide a framework for the assessment of both the distribution and diversification of a further set of non-ubiquitous 'accessory' genes associated with staphylococcal virulence and will also be a powerful tool in the future on which to assess gene content in staphylococcal lineages from micro array data. It will be utilised in this thesis to analyse the distribution of *sdrE* alleles within the natural population of *S. aureus* with the further aim of characterising the variation at this locus and elucidating the nature of evolution at this 'accessory' locus.

Revealing of the mode of evolution at virulence-associated loci may be important in understanding virulence potential in *S. aureus* isolates. The genome can been considered in two parts comprising 'core' and 'accessory' loci interspersed with lost and foreign genes (Lindsay, 2004). The final study in this thesis examines evolutionary parameters in both 'core' and 'accessory' loci to address whether recombination in the 'accessory' surface-exposed clumping factor B is responsible for the incongruence found within the 'core' MLST housekeeping gene.

CHAPTER TWO

MATERIALS AND METHODS

2.1 Bacterial Strains

Staphylococcus aureus strains used in this study have been isolated from the Oxfordshire and Nottingham, United Kingdom.

The Oxfordshire strains were isolated over a two year period from the anterior nares of asymptomatic blood donors and from patients with disease in hospital, and in the community. These have been previously characterised by MLST (Feil *et al.*, 2003).

Strains from Nottingham were isolated between October and November 2000 from asymptomatic nursing home residents and were kindly donated by Dr. Hajo Grundmann (Grundmann *et al.*, 2002).

Collection	isolation source	number
Oxford	hospital acquired disease	91
Oxford	community acquired disease	60
	total disease isolates	**151**
Oxford	asymptomatic carriage	174
Nottingham	asymptomatic carriage	160
	total carriage isolates	**334**

Table 1. Summary of bacterial strains.

Additional strains:

For platelet aggregation experiments

Lactococcus lactis strain MG1363 (Gasson, 1983) and *Staphylococcus aureus* strain Newman (Duthie & Lorenz, 1952) were used in platelet aggregation experiments courtesy of Professor Timothy Foster and Dr Louise O' Brien.

For representative population sample of *S. aureus*

Epidemic methicillin-resistant *S. aureus* (EMRSA) type strains: EMRSA3, EMRSA4 and EMRSA9 were kindly donated by Dr. Mark Enright (Enright *et al.*, 2000)

2.2 Preparation and Storage of Cell and DNA Stocks

S. aureus strains were grown on TS agar (Oxoid, UK) with 5% w/v defibrinated sheep's blood (TCS Microbiology, UK) for a minimum of 12hrs at 37°C.

Several colonies were resuspended in 1ml TS broth (Oxoid, UK) with 15% w/v glycerol and stored at -80°C providing frozen cell stocks.

Further colonies were resuspended in 400µl of lysis solution containing: 20µl Lysostaphin (500units ml-1), 20µl Lysozyme (5000units ml-1), 360µl TE Buffer (0.2ml 0.5M EDTA ph8.0, 1ml 1M Tris ph8.0, 98.8ml dH_2O).

Genomic DNA was extracted using the DNeasy tissue kit (Qiagen, Crawley, UK). DNA stocks were then stored at -20°C.

All stocks were catalogued and thermostably labelled for traceability and accurate identification.

2.3 Methods for Polymerase Chain Reaction and Sequencing

DNA amplification: Polymerase Chain Reaction

50µl PCR reactions were run in 46 well pink PCR plates (Abgene, UK) and contained:

36.3µl distilled H_20,

5µl 10X (supplied with Taq Polymerase)
5µl $MgCl_2$ (15mM supplied with Taq Polymerase)
0.2µl Taq polymerase (5u/µl: Promega, Wisconsin, USA)

1µl dNTPs (10mM each, Promega, Wisconsin, USA)
1µl forward primer (10pmol^{-1}, MWG Biotech)
1µl reverse primer (10pmol^{-1}, MWG Biotech)

0.5µl DNA

The following standard PCR conditions were use for all PCR reactions, unless otherwise stated. Annealing temperatures for each reaction were adjusted according to primer pairs.

PCR conditions:

- 3.00 min initial denaturation at 94°C,

- 34 cycles: denaturation at 95^{oC} for 30 secs
annealing (primer specific temperature) for 30 secs
extension at 72^{oC} for 1 min

- 10.00 min final extension step at 72^{oC}

All primer sequences can be found in the appendices as indicated in individual Chapters.

Electrophoresis of PCR product

The presence of amplicons was detected by gel electrophoresis stained with ethidium bromide.

-Agarose gel

A 1% agarose gel (10ml TBE Buffer 10X (Promega, Wisconsin, USA), 90ml dH2O, 0.8g agarose, 1µl Ethidium Bromide (10mg/ml) was run in 1% TBE buffer (100ml TBE Buffer 10X (Promega, Wisconsin, USA), 900ml dH_2O).

-Loading sample

5µl dH_20, 2µl loading dye (blue/orange 6X, Promega, Wisconsin, USA) and 5µl PCR product were mixed and loaded into the wells of the agarose gel. The samples were electrophoresed for 20-30 minutes at 150-180 volts and visualised the on the "Uvidoc" UV transilluminator.

Purification of PCR products

Amplicons were precipitated by the addition of 52µl of NaOAc/EtOH solution to PCR wells (2µl 3M NaOAc, 50µl 95% EtOH). The PCR plates were vortexed, sealed, centrifuged to at least 1000rpm for 10 seconds and then incubated at $-20^{\circ C}$ for 1 hour. After incubation they were centrifuged at 3500rpm, $4^{\circ C}$ for 1 hr. Immediately after the completion of the centrifuging, plates were inverted onto fresh blue roll and spun at 500rpm for 1 min in order to remove residual ethanol from the wells. The amplicon was washed by the addition of 150µl of 70% ethanol. The plates were resealed and spun at 3500rpm for 30 mins. The supernatant was discarded by inversion onto blue roll and centrifuging (still inverted upon fresh blue roll) dried the pellets at 750rpm for 1 minute. The purified amplicon was then resuspended in 15µl dH_20. Where small numbers of PCR products required purification the QIAquick PCR purification kit was used (Qiagen, Crawley, UK).

Sequencing of purified amplicons

Sequencing reactions contained 1µl of primer (1 $pmol^{-1}$), forward or reverse, 2µl Taq FS - Big Dye (Version 3, Applied Biosystems) and 2µl purified PCR product was placed in each

well (2 wells per isolate - forward/reverse). The plate was then sealed and spun to at least 1000rpm and held for 10 secs. The plates were then placed on the DNA engine. The sequencing PCRs were performed on with an initial 10 second denaturation at 96°C followed by 24 cycles at 50°C, extension at 60°C for 2 mins and then 4°C forever.

Precipitation of sequence cycling products

The final precipitation of the sequence cycling products was done by the addition of 52µl of NaOAc/EtOH solution (2µl 3M NaOAc, 50µl 95% EtOH). Plates were once again vortexed, sealed, centrifuged to at least 1000rpm for 10 seconds and then incubated at –20°C for 1 hour. After incubation they were centrifuged at 3500rpm, 4°C for 1 hr. Immediately after the completion of the centrifuging, plates were inverted onto fresh blue roll and spun at 500rpm for 1 min in order to remove residual ethanol from the wells. The DNA pellet was washed by the addition of 150µl of 70% ethanol. The plates were resealed and spun at 3500rpm for 30 mins. The supernatant was discarded by inversion onto blue roll and centrifuging (still inverted upon fresh blue roll) dried the pellets at 750rpm for 1 minute. The plates were then sealed and stored at –20°C until ready for loading on the ABI Prism 3700 DNA sequencer.

2.4 Cloning and Expression of *S. aureus* surface proteins

2.4.1 Expression Vector pkS80

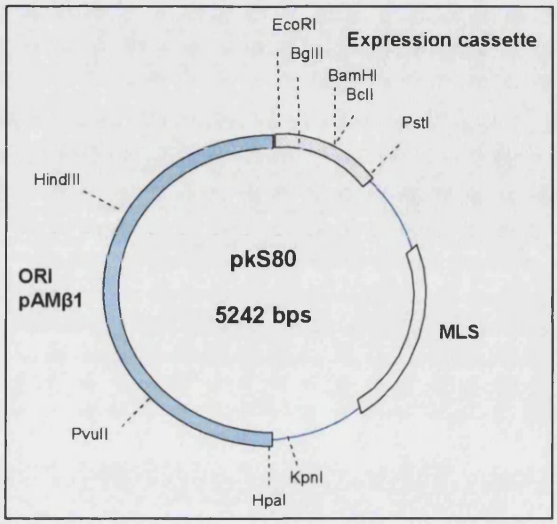


Figure 1. Map of the pKS80 plasmid. The positioning of the plasmid replication origin of pAMβ1 and the macrolide, lincosamide and streptogramin B (MLS) marker are indicated in bold. The expression cassette is shown in blue lines and is shown in detail below in Figure 2.

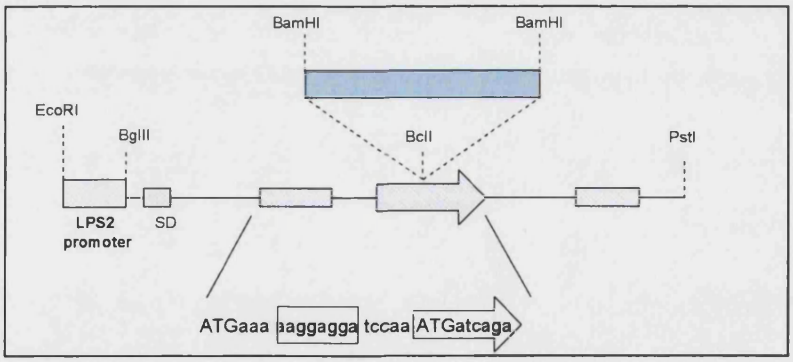


Figure 2 . The pkS80 expression cassette

Expression of the target gene is achieved by cloning into the BclI site TGATCA within the large open arrow. Translation is optimised by fusing the target open reading frame

ATG codon to the ATG codon (capitals) of the expression cassette. This overlaps the stop codon TGA of the upstream open reading frame. The ribosome-binding site is indicated by the upstream box. The lactococcal bacteriophage C2 promoter LPS2 provides constitutive expression of a short open reading frame and its positioning with the Shine-Dalgarno (SD) sequence is shown in figure 2.

2.4.2 Preparation of vector and insert

sdrE gene amplification

sdrE genes were amplified in a 100µl reaction with PfU polymerase (Promega, UK) containing:

80µl distilled H_20,

10µl 10X with $MgSO_4$ (supplied with PfU Polymerase)
2µl PfU polymerase (xu/µl: Promega, Wisconsin, USA)

2µl dNTPs (10mM each, Promega, Wisconsin, USA)
2µl forward primer (100pmol^{-1}, table 2)
2µl reverse primer (100pmol^{-1}, table 2)

2µl DNA

Primers	Primer sequence 5'- 3'
sdrE forward	CCGGGATCCTGATTAACAGGGATAATAAAAAG
sdrE reverse	CCGGGATCCTTATTTGTTTTGTTTTTTGCGACG

Table 2. *SdrE* gene primers. The BamHI restriction site is underlined.

PCR conditions:

- 1.00 min initial denaturation at 94°C,

- 30 cycles: denaturation at 95°C for 30 secs
annealing 58°C for 30 secs
extension at 72°C for 7 min

- 10.00 min final extension step at 72°C

The PCR product was then prepared for digestion using the High Pure PCR product purification kit (Roche Diagnostics, Switzerland).

Digest of *sdrE* gene

The *sdrE* gene was digested in a 100µl reaction and incubated at 37°C in a water bath overnight.

BamHI enzyme	(New England Biolabs, UK)	2µl
10X Buffer B	(supplied with enzyme)	10µl
Purified *sdrE* product		50µl
distilled H_20,		38µl

The digested *sdrE* gene was then prepared for ligation using the High Pure PCR product purification kit (Roche Diagnostics, Switzerland).

Preparation of vector: Digest of pkS80

50 µl digestion of pKS80 vector was incubated at 37°C overnight.

BclI enzyme	(New England Biolabs, UK)	1µl
10X Buffer M	(supplied with enzyme)	5µl
pKS80		10µl
dH20		34µl

Dephosphorylation of cut plasmid

This step is necessary to promote the attachment of the sdrE insert to the pkS80 vector sequence rather than back onto itself.

Digest	50µl
10X Buffer	5µl (5µl of buffer in digest)
Alkaline phosphatase (New England Biolabs, UK)	1µl
dH20	44µl

The 100 µl dephosphorylation reaction was incubated at 37°C for 1 hour. The reaction was stopped with 2µl 0.5M EDTA and incubated at 65°C for a further 10 mins before continuing directly to the purification of the dephosphorylated plasmid using the High Pure PCR product purification kit (Roche Diagnostics, Switzerland). The plasmid was then ready to be stored at -20°C.

2.4.3 Ligation of vector and insert

Ligation

pKS80	8µl
insert	8µl
10X buffer	3µl
T4 DNA ligase (Promega, WI, USA)	2µl
dH20	9µl

The ligation reactions were incubated overnight at 14°C.

Ethanol precipitation of ligation product (it is essential that no salt remains)

1. The post ligation reaction was made up to 50µl with dH_20.

2. To the reaction was added: 3M NaOAc 5µl
Ice cold absolute EtOH 200µl

3. Reaction tubes were then incubated at -80°C for 30mins.

4. Spun at 13,000 rpm for 10 minutes.

5. The supernatant was removed.

6. 100µl of 70% (v/v) EtOH was added.

7. Spun at 13,000 rpm for 10 minutes.

8.The supernatant was removed and incubated at 37°C to evaporate excess EtOH.

9. The pellet was resuspended in 5µl of dH_20.

2.4.4 Transformation of *Lactococcus lactis*

Buffers for electroporation of electro competent *Lactococcus lactis*

Storage Buffer

0.5M sucrose

10% (v/v) Glycerol

GM17 containing 2.5% glycine

200ml M17 broth (Difco, USA)

0.5% (w/v) glucose

2.5% (w/v) glycine

Bottle top filter to maintain sterility.

Starter culture broth

100ml M17 broth

0.5% (w/v) glucose

SGM17MC Recovery broth 20ml

GM17 broth - double strength

0.5M sucrose

20mM $MgCl_2$

2mM $CaCl_2$

Erythromycin Selective plates

200ml M17 agar (Difco, USA)

0.5% (w/v) glucose

Erythromycin 5μg/ml (Sigma, UK)

Preparation of electro competent *Lactococcus lactis*

1. *L. lactis* was grown in M17 containing 0.5% (v/v) glucose at 30°C for 16 hours (starter culture).

2. GM17 containing 2.5% (w/v) glycine was inoculated from starter culture.

3. Cells were grown until they reached an OD of 0.5-0.6 at A600nm.
(optimum competency in growth phase)

4. The culture was chilled on ice for 10 minutes.

5. Cells were harvested by centrifugation in a cooled rotor.

6. Cells were resuspended in 1/10 volume (20ml) of ice cold storage buffer.

7. Cells were washed again in ice cold storage buffer.

8. Cells were recovered again by centrifugation.

9. Cells were then resuspended in 1/100 (2ml) volume of ice cold storage buffer.

L. lactis electroporation and transformation

1.100µl *L.lactis* cells were added to the 5µl ligation previously prepared (Step 9, page 50).

2. Cells were carefully transferred to an ice cold 0.2cm electroporation cuvette and tapped to the bottom of the cuvette.

3. The cuvette was placed in the electroporation chamber until the cuvette was seated between the contacts in the apparatus.

4. One pulse was used: 2.5kV field strength
25µF capacitance
400 ohms resistance

producing a pulse time of about 4.8msec.

14. 0.96ml of SGM17MC recovery broth was immediately added to the cells which were then transferred to a microfuge tube.

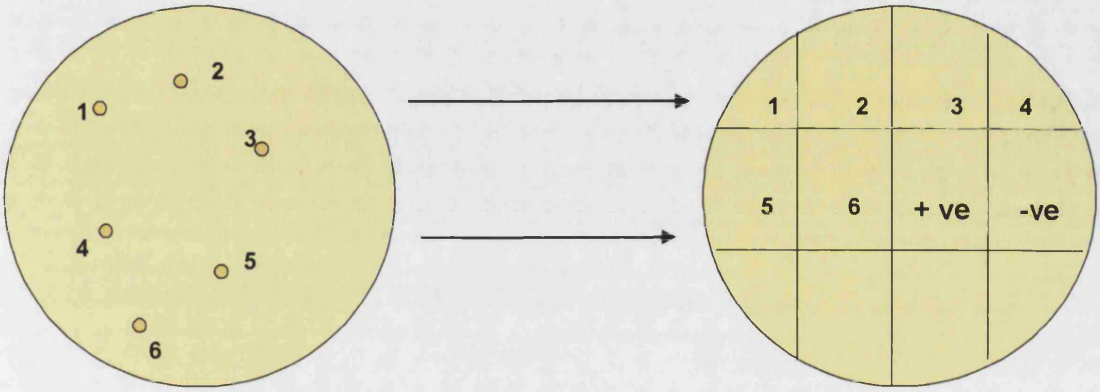
15. Cells were incubated on ice for 10 minutes.

16. Cells were then incubated at 30°C for 2 hours in a waterbath.

17. Cells were plated on GM17 agar with erythromycin.

18. Plates were left at 30°C for 2 days to allow colonies to grow.

19. Single colonies were picked and touched onto marked plate for traceability (see below) and incubated at 30°C overnight.



20. Loops were then agitated in 2ml culture to resuspend the remaining cells.

21. Control cultures were also prepared:

positive – *L.lactis* with *sdrE* insert (from *S. aureus* Newman)
negative – Wild type *L.lactis* (without insert)

2.4.5 Analysis of Surface proteins

Manipulation of colonies for screening assay

The presence of an erythromycin marker means that only *L.lactis* cells containing the pkS80 plasmid are able to grow on this medium. Although this represents successful transformation, we assume successful original ligations of plasmid and insert and subsequent expression of the sdrE protein. Successful expression can be determined by screening for the sdrE protein.

1. 10µl of each culture, prepared above, to was transferred to marked nitrocellulose transfer membrane (PROTRAN, Schleicher and Schuell).

2. The culture was allowed to spread and dry on the membrane.

3. When the membrane had dried it was transferred to an empty tip box with tweezers to avoid contamination.

4. The membrane was covered with 50ml 10% (w/v) dried skimmed milk (Marvel) solution in TBS buffer (29.24g NaCl, 2.42g Tris, 1L dH_20).

5. This was shaken for 1.5 h on a benchtop shaker.

6. The milk solution was removed.

7. 10ml 10% (w/v) milk solution with 10µl primary antibody (rabbit anti-sdrE A region antibodies kindly donated by T.J.Foster; 1:1000) was poured over the membrane.

8. The immersed membrane was shaken for 1.5 h on a benchtop shaker.

9. The milk solution was removed and the membrane washed in TBS buffer). Twice.

10. The membrane was then left covered in TBS buffer, shaking, for a further 10 minutes.

11. Steps 9 and 10 were repeated 3 times.

12. The membrane was then immersed in 10ml 10% (w/v) Marvel solution with 5µl protein A coupled to horse radish peroxidase (Sigma, UK) (1:2000).

13. The membrane was left for 1 h at room temperature.

14. Washing in steps 9 and 10 was repeated 3 times.

15. The membrane stored in dH_20 during preparation for the final stage.

16. Colonies with sdrE expression on the membrane were recognised by chemiluminesence using the LumiGLO system (KPL).

Preparation of surface proteins for SDS page

1. Equal volumes of culture were centrifuged at 3000rpm for 10 mins.

2. The supernatant was discarded and the remaining fluid aspirated.

3. The pellet was resuspended in 25ml of 1 x PBS.

4. The optical density (OD) at A600nm was recorded.

5. Appropriate volumes for an OD of 10 at A600nm were aliquoted.

6. Tubes were spun at 10,000 rpm for 2 minutes.

7. The supernatant was removed from each.

8. Each pellet was resuspended in:

30% (w/v) Raffinose	250µl	(Sigma, UK)
Mutanolysin	50µl	(Sigma, UK)
Lysozyme	10µl	(Sigma, UK
Proteinase inhibitors (complete mini tablet)	8µl	(Roche, UK)

9. The tubes were then incubated at 37°C for 20 minutes, inverting every 5 minutes.

10. Tubes were then spun at 8000rpm for 10 minutes.

11. The supernatant was removed and retained.

Preparation of SDS page gels

8% Resolving gel

30% (w/v) acrylamide mix	1.3ml
1.5M Tris (pH8.8)	1.3ml
10% (v/v) SDS	50µl
10% (w/v) ammonium persulfate	50µl
TEMED	3µl
H_2O	2.3ml

5% Stacking Gel

30% (w/v) acrylamide mix	170µl
1.5M Tris (pH6.8)	130µl
10% (v/v) SDS	10µl
10% (w/v) ammonium persulfate	10µl
TEMED	1µl
H_2O	680µl

1. Resolving and stacking gels were cast.

2. Wells were loaded.

Loading sample: 100µl supernatant of surface proteins (prepared step7 page x)
100µl sample buffer (Laemmli, 1970)

With running conditions: Stacking gel 20 mins at 80V
Resolving gel 90 mins at 120V

3. The western membrane with transferred proteins was blocked with 10% milk solution as before and left overnight.

4. Milk solution was removed and replaced with 10ml milk solution with 10µl primary antibody.

5. This was incubated for 1.5 h.

6. The membrane was washed several times with 1x TBS buffer with 0.1% (v/v) Tween.

7. 10ml milk solution with 5µl protein A coupled to horse radish peroxidase (Sigma, UK) was poured over the membrane.

8. The membrane was washed several times with 1x TBS buffer with 0.1% (v/v) Tween and stored in dH_20 during the preparation for the final step.

9. Bound antibodies were recognised by chemiluminesence using the LumiGLO system (KPL, Maryland, USA).

2.4.6 Platelet aggregation experiments

Preparation of transformed *L. lactis* cells for aggregation

1. Equal volumes of cells (from where) were centrifuged at 3000rpm for 10 minutes.

2. The supernatant was discarded and aspirated.

3. The pellet was resuspended in equal volumes of 1 x PBS.

4. The supernatant was discarded and aspirated.

5. The pellet was resuspended in 1ml PBS.

6. ODs at A600nm were recorded.

7. Cultures were prepared to an OD of 1.6 at A600nm in a total volume of 1ml PBS. This is the optimum cell concentration for aggregation experiments.

Platelet preparation

1. Nine volumes of blood were drawn from two healthy volunteers by venepuncture and minimum stasis into one volume of 3.8% (w/v) sodium citrate.

2. Platelet – rich plasma (PRP) was prepared by centrifugation of anticoagulated whole blood at room temperature at 150g for 10 minutes.

3. The supernatant (PRP) was removed and kept for aggregation experiments.

Platelet aggregation

1. 50µl of transformed *L.lactis* cells were added to 450µl of PRP.

2. Platelet aggregation was assayed by light transmission at 37°C using a PAP-4 aggregometer (Bio/Data, PA, USA).

2.5 Nucleotide sequence analysis

2.5.1 Nucleotide sequence assembly

Nucleotide sequence trace files were assembled and edited using SEQMAN, DNAstar (Lasergene Inc, WI, USA). The coding strand was established using the translation tool in EDITSEQ, DNAstar (Lasergene Inc, WI, USA). ClustalX (Thompson *et al.*, 1997) was used for the multiple sequence alignment. Fragment size and conserved start and stop sequences for genes and gene fragments were identified for individual genes using ClustalX

2.5.2 Phylogenetic reconstruction

Neighbour-joining method (Saitou & Nei, 1987)

Neighbour-joining tree created using ClustalX (Thompson *et al.*, 1997). This is a widely used distance method for tree building. Distance methods are based on the idea that if we knew the actual evolutionary distance between all members of a set of sequences, then we could easily reconstruct the evolutionary history of those sequences. A distance matrix is created in which a pair of sequences which have the shortest distances between them are defined as 'neighbours' and form a new node. The distance matrix is updated and a nearest distance neighbour to this new node is then identified. This additive process continues until all sequences are included. Neighbour-joining does not optimise a criterion of fit between the tree and the data. However it is a fast clustering method and is suitable to illustrate relationships between protein sequences as it is used in this thesis.

Maximum Likelihood (ML)

Maximum Likelihood trees were generated using PAUP* version 4.0b10 (Swofford, 2000). The maximum likelihood method of phylogenetic reconstruction is a discrete method which uses the sequence data directly, in contrast to distance methods which infer phylogeny through the conversion of sequence data to pairwise distances. Maximum likelihood generates a tree (a neighbour-joining tree has been used in this thesis) and searches to improve it, outputting an ML tree it considers the most likely given the observed data. The maximum likelihood estimate should not be misinterpreted as the probability that the tree is the true tree, but rather it is the probability that the tree has given rise to the observed data. The likelihood estimate for any given full tree is the sum of the likelihoods for all sites calculated separately. To calculate the likelihood, all evolutionary scenarios for the presence of a particular nucleotide at any given site must be considered. So the likelihood is the summation of the probabilities of every possible reconstruction of ancestral states, given a model of base substitution. In this way the evolutionary model is important as this will influence individual site likelihoods and thus the full tree likelihood. Maximum likelihood has several advantages over other methods. It evaluates different tree topologies based upon all the sequence information. The result is dependent upon the model of evolution used, however it is not necessary to assume a model, an appropriate model can be estimated from the data. The HKY85 model of nucleotide substitution has been used for tree reconstruction in this thesis, with the transition/transversion (Ti/Tv) ratio, base frequencies and gamma shape parameter optimized. (The gamma parameter describes the extent of nucleotide substitution rate variation between sites assuming a discrete gamma distribution with 4 categories). So although it is computationally intensive and thus slow, it is a popular method as it allows the incorporation of explicit models of sequence evolution and also permits statistical tests of evolutionary hypotheses.

Bayesian Reconstruction

Implemented using MrBayes ver 2.01 (Huelsenbeck & Ronquist, 2001)

There is a subtle and yet fundamental difference between Maximum Likelihood and Bayesian analysis;

-Maximum likelihood result: the probability of the data given the tree
-Bayesian result: is expressed as the probability of the tree given the data

Both use the likelihood function but whereas free parameters can be optimised for maximum likelihood, thus maximising the likelihood, in Bayesian analyses the *posterior probability density* of the tree topologies and model parameters are sampled. The *posterior probability density* is approximated by the *Markov chain Monte Carlo* (MCMC) method. MCMC begins at a given point in 'tree-space' this could be a user defined tree, such as one generated by neighbour-joining, or a random tree. This tree will have a likelihood associated with it but not the optimised, maximum likelihood. A new state is randomly proposed and if it has a better likelihood the chain accepts this tree. If the likelihood of the proposed state is only a little worse, it may be accepted and the chain is able to cross likelihood valleys. In this way Bayesian analysis has advantages over a hill-climbing algorithm such as maximum likelihood. The analogy can be thought of like this: in trying to get to the highest point on the earth, you would climb a mountain. Imagine you were blindfolded. In order to get to the top you take single steps, if the ground is lower you would go back and take a new step until you reached higher ground. Eventually there comes a point on the mountain when the ground is lower in all directions. This would be considered the highest point of the mountain, but is it the highest point on the earth? In order to determine this, you would have to sample more mountains. Maximum likelihood climbs only one mountain whereas Bayesian analysis as implented in MrBayes can climb many thus exploring a greater range of 'tree-space'. This is also enhanced with the introduction of Metropolis-coupled MCMC where several chains run in parallel (Huelsenbeck & Ronquist, 2001). All but one of the chains is 'heated' whereby the acceptance probabilities are increased allowing easier crossing of the likelihood valleys. Chains are allowed to swap with the cold chain but only the cold chain is sampled. The MCMC is sampled every 10 or 100 generations. The chain will run for many generations with the likelihood increasing steadily until it eventually reaches equilibrium. At this point MrBayes samples trees according to their posterior probabilities. Trees sampled after equilibrium is reached can then be used to generate a consensus tree. Trees sampled before equilibrium are discarded as 'burn-in'.

However, Bayesian analysis is computationally intensive and it is unclear how long a chain should be left to run. Jumps in the likelihood, long after apparent equilibrium has been reached, have occasionally been observed. This problem may be eliminated given sufficient "burn in" time. Statistical confidence of interior branches is generally higher and has generated conflict in ideas whether bootstrap support is too conservative or posterior probabilities are too liberal (Misawa & Nei, 2003; Suzuki *et al.*, 2002). However, posterior probability is a more intuitive measure compared to bootstrapping which is a popular method for estimating sampling error.

Splits Decomposition

Splits Decomposition has been implemented in SplitsTree (Huson, 1998).

Since evolutionary data can often contain a number of different, and sometimes conflicting, phylogenetic signals it does not always clearly support a unique tree. Bandelt and Dress addressed this issue with the development of the Splits decomposition method. For ideal data this will produce a tree, whereas in the presence of conflicts it will appear network-like and can be interpreted as possible evidence for different and conflicting phylogenies within the dataset (Bandelt & Dress, 1992).

2.5.3 Protein homology modelling

Homology modelling was done using the program MOE (Molecular Operating Environment; Chemical Computing group, Montreal). The *S. epidermidis* SdrG crystal structure (Protein Data Bank accession number 1R17.pdb) was used as the template for constructing SdrE homology models (amino acid sequence identity is 49%: allelic variants sdrE and bbp share only 63% in this region) Ten models were generated and subjected to a coarse energy minimisation for the target protein. The best model with the highest residue packing scores was selected for further full energy minimisations using a CHARMM22 force field distributed in MOE. Ribbon diagrams of the SdrE homology model were produced with MOLSCRIPT (Kraulis, 1991) and rendered with POV-RAY™.

2.5.4 Evidence for recombination from nucleotide polymorphisms

Sawyer's Runs Test

The Sawyer's Runs Test is a modified alternative to Stephens original model (Stephens, 1985) for the detection of statistically significant block structure within a set of sequences. Both tests look for evidence of recombinational exchanges within a set of aligned sequences by determining if regions of sequence pairs have more consecutive identical polymorphic sites in common than would be expected by chance (Drouin *et al.*, 1999; Sawyer, 1989). In Stephens test the dataset is partitioned into only two subsets, the partition of which can be difficult to appropriately identify if there are many moderately to highly polymorphic sequences. In this way Sawyer's Runs Test is more appropriate for larger datasets. Pair-wise comparisons of derived sequences containing only silent polymorphic sites (condensed sequences) are made. Each pair of sequences is then partitioned into fragments containing runs of identical sites. The lengths of all the fragments found between every pair-wise comparison are used to obtain two values:

the sum of the squares of condensed fragments (SSCF)

A condensed fragment being the number of polymorphic sites which lie linked in a 'run' between two discordant sites or one discordant site and the end of the sequence.

the maximum condensed fragment (MCF).

This being the largest of all such fragments and for all pairs

This method assumes a constant mutation rate across sequences, but by using silent polymorphic sites the SSCF and MCF are not significantly influenced by mutational hot and cold spots. A gene conversion event is likely to increase the values of SSCF and MCF as it results in an identical region within two sequences and may produce an unusually long fragment. The Sawyer's Runs Test generates artificial data sets based on a user defined number of site permutations of >10,000. These permutations rearrange the sites which are assigned a class based on degeneracy. Permutation of the sites is constrained so that sites of a given class can only be assigned to site positions of the same class in the original dataset. SSCF and MCF values are calculated. Thus, in the presence of recombination the

SSCF and MCF values for the observed sequences are expected to be greater than those obtained for the randomly permuted sequences. The significance of the difference between the value for observed and permuted sequences is described by the p values. Alternative measures of the sizes of conserved segments are obtained by performing the tests on all nucleotide sites (uncondensed sequences). For each pair-wise partition, fragment boundaries are defined by the presence of a discordant silent polymorphic site. The corresponding uncondensed fragment statistics, SSUF and MUF, are more likely to resolve large conversion events but do not control for mutational hot and cold spots.

However, both the Stephens original model and Sawyer's Runs Test have the disadvantage of not being able to define the limits of blocks.

Download Sawyer's Runs Test
(http://www.biols.susx.ac.uk/Biochem/Molbiol/sawyers-runs-tests.html)

Maximum chi-squared test

The Maximum-chi squared test was developed to enable the identification of recombinational replacement boundaries based on the maximum chi squared method by John Maynard Smith (Smith, 1992). This program compares the distribution of polymorphic sites along two aligned sequences, or two "donor" sequences and a probable "recombinant". Significant clustering of polymorphic sites is evident if the observed clustering of polymorphic sites is greater than that obtained in randomly generated datasets containing the same number of polymorphic sites. A non random distribution of polymorphic sites along two sequences may be due to a recombinational event. The program also finds the putative recombination boundary or crossover point based on the most significant partioning of the sequences. Despite this the application of this method is limited. It is most efficient when only those sequences which have been involved in the recombination event are considered, namely the two parental molecules and the putative mosaic. The inclusion of polymorphic sites from other sequences may lead to the identification of false boundaries.

Population-scaled recombination rate (ϱ)

The key parameter in determining the extent of linkage disequilibrium is the product of the recombination rate and effective population size, this is the population-scaled recombination rate (ϱ) $N_e r$. The pairwise program of the LDhat package by Gil McVean, estimates this parameter using the approximate likelihood coalescent method of Hudson (Hudson, 2001). For each pair of segregating sites the coalescent likelihood of observing the data under a range of population recombination rates is estimated by the importance sampling method of Fearnhead and Donnelly (Fearnhead & Donnelly, 2001). The likelihoods are combined across pairs to provide a point estimate of the population recombination rate.

LDhat available at www.stats.ox.ac.uk/~mcvean/LDhat.html

Minimum number of recombination events (R_M)

R_M denotes the minimum number of recombination events in the history of the sample (thus R_M will underestimate the total number of recombination events) implied by the data using the four-gamete test. The sample size must be at least four and only segregating sites are compared. Hudson 1983 developed a method for simulating samples from the neutral infinite-site model with finite recombination. The history of a sample is a collection of correlated family trees, and one is generated for each site. The family tree for a site traces the genealogy of a site back to its most recent common ancestor indicating which sampled gametes are most closely related and when the most recent common ancestors occurred (Hudson & Kaplan, 1985).

2.5.5 Recombination and phylogenetic consistency

Bellerophon: detecting chimeric sequences

Bellerophon was specifically developed to detect 16S rRNA gene chimeras in PCR-clone libraries. In the same way it can be used to detect mosaic sequences within a gene dataset. Bellerophon detects chimeras based on a partial treeing approach (Hugenholtzt & Huber, 2003; Wang & Wang, 1997). In this way phylogenetic trees are inferred from independent regions (fragments) of a multiple sequence alignment and the branching patterns are

compared for incongruencies that may be indicative of a chimeric or mosaic sequence. This method compares topologies of phylogenetic trees inferred from independent regions (fragments) of a sequence alignment. Incongruencies in branching patterns may be indicative of chimeric sequence. (Huber *et al.*, 2004). The limitation of this approach however, is that only chimeric (mosaic) sequences are identified where there is are putative parental sequences within the dataset. For smaller sequences there is also the problem of window size. Typically only two windows can be used. This program has the advantage of accepting large aligned sequence datasets within which putative mosaic and parental sequences are identified. This provides a useful first step in the utilisation of the maximum chi squared test when using large numbers of sequences.

Congruence analysis

Phylogenetic congruence analysis can be used as a measure of phylogenetic consistency and indicative of recombination. It is expected that in the absence of recombination unlinked loci would have a shared evolutionary history. However, extensive recombination can obliterate the shared phylogenetic signal so phylogenies from genes in the same genome are no longer phylogenetically consistent with each other, or no more similar to each other than they are to trees of random topology. This notion is the cornerstone for the maximum likelihood method which examines whether the a tree generated from one gene is a significantly better fit to the data from another, than trees of random topology (Feil *et al.*, 2001). The approach compares the maximum likelihood scores of gene trees against the 99th percentile of the distribution of the scores for trees of random topology (200 used in this study) given the reference data. Two genes, A and B. are scored as significantly congruent if the difference between the likelihood scores of the trees for gene A and gene B (Δ-*lnL*) is lower than the Δ-*lnL* between the 99th percentile of the scores of the random trees and the likelihood score for gene A.

(Likelihood values for trees are expressed as log likelihood (-*lnL*) since the individual likelihoods are extremely small numbers).

This method has been used to compare the phylogenetic congruence between MLST genes in a number of bacterial species (Feil *et al.*, 2001), including *S. aureus*.

Maximum likelihood trees generated as described in phylogenetic reconstruction were used for congruence analysis. The likelihood scores for trees of random topology were also computed using the same model and parameter optimisation. The generation of maximum likelihood trees as described previously, random trees and maximum likelihood scoring was calculated in PAUP* 4.0b10 (Swofford, 2000).

Comparisons of phylogenetic reliability

Resolution in a concatenated sequence tree is provided by informative sites within each of the genes included. The number of informative sites varies quite dramatically for genes within the dataset so each gene will be weighted differently, ie contribute varying amounts of information. This means that genes with more informative sites have more weight in the tree than those with fewer sites. In order to independently score the topology of any individual loci against the topology of all the remaining data, that loci must be removed from the dataset and be absent from the concatenated data tree it is to be scored against. Each time the dataset is modified, model parameters will change with reoptimisation so ML scores will not be comparable between tests.

The Shimodaira-Hasegawa (SH) test calculates the difference in likelihood score from the highest scoring tree (which will invariably be the tree generated from the test data) to all other genes individually. It also measures the significance of this difference (Shimodaira and Hasegawa, 1999). Using the difference in likelihood score from the best tree in each instance can be considered a comparable way of ranking the phylogenetic reliability of genes. The SH test was implemented in PAUP* 4.0b10 (Swofford, 2000).

2.5.6 Testing for selection

The degeneracy of the genetic code ensures that synonymous substitutions (d_S) do not change the encoded amino acid. However, nonsynonymous substitutions (d_N) change the amino acid sequence in the encoded protein.

Tests for positive selection are based upon the idea that synonymous mutations are selectively neutral and can become fixed in a population primarily as a result of stochastic

mechanisms such as drift. If all nonsynonymous substitutions were neutral, then their rate of occurrence per site (d_N) would be equal to that of synonymous substitutions (d_S), so that d_S/d_N equals one. A lower ratio of synonymous to nonsynonymous substitutions per site ($d_S/d_N > 1$) means that some proportion of the nonsynonymous mutations are deleterious and removed by purifying selection. Conversely positive selection fixes advantageous nonsynonymous substitutions faster than genetic drift fixes synonymous mutations resulting in a $d_S/d_N < 1$. Ratios of synonymous and nonsynonymous substitutions (d_S/d_N) over single loci fragments have been calculated in MEGA version 2.1 (Kumar *et al.*, 2001).

Datamonkey was used to identify individual sites within single loci under selective pressures (available online at www.datamonkey.org) (Kosakovsky Pond *et al.*, 2004).

2.5.7 Statistics

Chi-squared test of association

This is a non-parametric test of association between two variables for bivariate tabular analysis and so is used for categorical data. This tests whether the variable in the columns is contingent (independent) of variable in the rows. The chi-square test measures the difference between the observed and the expected values given the null hypothesis (that there is no relationship between row and column frequencies). This test requires that:

1. the sample is randomly drawn from the population
2. the data is in raw frequencies
3. the measured variables are independent
4. values are exclusive and exhaustive
5. the observed frequencies are not too small

The mathematical equation for this statistical test is:

$\chi^2 = \sum (O - E)^2 / E$ (where O = observed frequencies E = expected frequencies)

The value for chi square is also known as the 'critical value'. The degree of freedom is calculated by subtracting one from the total number of categories in the study.

Where P= 0.01 the chance of the observed distribution being due to chance is 1 in 100 and you are 99% percent certain the null hypothesis is wrong.

Analysis of variance: ANOVA and the Student's t-test

An ANOVA is a parametric test closely related to the t-test. The major difference is that, where the t-test measures the difference between the means of two groups, an ANOVA tests the difference between the means of two or more groups. Determining the statistical difference between means is calculated through the comparison of variances. The advantage of an ANOVA over multiple t-tests is that it reduces the probability of Type I errors. The drawback however with an ANOVA is that you lose specificity. All the result will tell you is whether they are all significantly different from each other, not which groups are significantly different from each other. An ANOVA may not be appropriate if there is strong reason to doubt the assumption of equal variances between groups.

Analysis of variance: Kruskal-Wallis test

This is a non-parametric alternative to an ANOVA which can be used when there is a high level of variation between groups. This test determines whether there is a difference between groups by calculating whether one group tends to have higher ranking values than comparison groups.

Regression analysis

Simple linear regression aims to find a linear relationship between a response variable and a possible predictor variable by the method of least squares. A regression line is drawn through the points on a scatter plot to summarise the relationship between the variables being studied. When it slopes down this indicates a negative relationship between the variables, when it slopes upwards a positive relationship is indicated. If we find evidence of relationships between variables we can use the technique of correlation to test the statistical significance of the association.

CHAPTER THREE

GENE FUNCTION, RECOMBINATION AND PHYLOGENY

CHAPTER THREE: GENE FUNCTION, RECOMBINATION AND PHYLOGENY

3.1 INTRODUCTION

This chapter discusses the use of *S. aureus* as a model for addressing fundamental questions regarding evolutionary parameters:

1. How does the functional category of genes affect selective constraint?

2. Do housekeeping genes represent a reliable and consistent sample of the 'core genome'?

3. Does the 'complexity hypothesis' apply at the intraspecific level conferring a measure of immunity to recombination within information pathway genes?

4. Is there a relationship between gene function and the frequency of recombination?

5. Is there a relationship between phylogenetic reliability and gene function?

Data from multiple loci will be compared to examine whether a robust intraspecific phylogeny can be reconstructed for *S. aureus*

Bacterial Strains

This study examined 30 strains, all of which have been previously characterised by MLST (Feil *et al.*, 2003). These are shown in Table 1 and represent 26 unique sequence types and include 6 MRSA strains, 3 of which are epidemic-MRSA (eMRSA). Two strains are included for some STs of particular clinical relevance. The majority of strains represent a diverse sample from a collection of 334 isolates representing hospital-acquired disease, community-acquired disease and asymptomatic carriage collected in the Oxfordshire region. As these strains were recovered from different epidemiological backgrounds (including asymptomatic carriage) they should represent a reasonable sample of a localised population at a particular point in time.

Strain	ST	History	Resistance Profile
H512	1	Hospital -acquired disease	MSSA
EMRSA3*	5	EMRSA type strain	MRSA
H466	5	Hospital -acquired disease	MSSA
C2	7	Community-acquired disease	MSSA
H591	8	Hospital -acquired disease	MSSA
H116	9	Hospital -acquired disease	MSSA
H19	10	Hospital -acquired disease	MSSA
H402	13	Hospital -acquired disease	MSSA
H783	15	Hospital -acquired disease	MSSA
D274	17	Asymptomatic carriage	MSSA
D17	20	Asymptomatic carriage	MSSA
C640	22	Community-acquired disease	MSSA
C720	22	Community-acquired disease	MRSA
C437	25	Community-acquired disease	MSSA
C101	30	Community-acquired disease	MSSA
H325	36	Hospital -acquired disease	MRSA
H831	36	Hospital -acquired disease	MRSA
H295	45	Hospital -acquired disease	MSSA
H707	49	Hospital -acquired disease	MSSA
H417	50	Hospital -acquired disease	MSSA
D97	55	Asymptomatic carriage	MSSA
D535	59	Asymptomatic carriage	MSSA
D547	97	Asymptomatic carriage	MSSA
D456	101	Asymptomatic carriage	MSSA
H560	121	Hospital -acquired disease	MSSA
D365	121	Asymptomatic carriage	MSSA
D22	182	Asymptomatic carriage	MSSA
D470	207	Asymptomatic carriage	MSSA
EMRSA4*	239	EMRSA type strain	MRSA
EMRSA9*	240	EMRSA type strain	MRSA

Table 1. Bacterial Strains - * kindly donated by Dr Mark Enright.

Choice and classification of loci

The genome sequencing paper for *S. aureus* strain N315 (Kuroda *et al.*, 2001) was used to select loci from diverse physical locations and from both leading and lagging strands around the genome. The functional classification of genes was adopted from the *Bacillus subtilis* study (Kunst *et al.*, 1997). There are 3 defined classes included in this study are: Informational Pathways: DNA replication, regulators; Housekeeping: Intermediary metabolism; Cell envelope and cellular processes: sensors, cell wall. Also included are undefined genes of unknown function and ORPHANS (no similarity to any genes of unknown function). Primers can be found in Appendix A1.

Gene	Locus*	Function**	Functional Category***
hutH	SA0008	histidine ammonia lyase	Housekeeping
serS	SA0009	seryl-tRNA synthetase	Information Pathways
SA0013	SA0013	conserved hypothetical protein	Unknown function
dnaC	SA0015	replicative DNA helicase	Information Pathways
vicK	SA0018	two component sensor hisidine kinase	Cellular envelope and cellular processes
SA0100	SA0100	conserved hypothetical protein	Unknown function
SA0117	SA0117	similar to rhizobactin siderophore biosynthesis protein	other function
SA0139	SA0139	hypothetical protein	ORPHANS
adhE	SA0143	alcohol-acetaldehyde dehydogenase	Housekeeping
hsdR	SA0189	probable type 1 restriction enzyme restriction chain	Information Pathways
SA0224	SA0224	similar to 3-hydroxyacyl- CoA dehydrogenase	Housekeeping
SA0268	SA0268	hypothetical protein	ORPHANS
SA0272	SA0272	hypothetical protein similar to transmembrane protein Tmp7	Cellular envelope and cellular processes
SA0275	SA0275	conserved hypothetical protein	Unknown function
yqiL	SA0342	acetyl-CoA acetyltransferase	Housekeeping - MLST
tufA	SA0506	translational elongation factor TU	Information Pathways
pta	SA0545	phosphate acetyltransferase	Housekeeping - MLST
sarA	SA0573	staphylococcal accessory regulator A	Information Pathways
tpi	SA0729	triosphosphate isomerase	Housekeeping - MLST
SA0740	SA0740	hypothetical protein	ORPHANS
SA0775	SA0775	conserved hypothetical protein	Unknown function
SA0778	SA0778	conserved hypothetical protein	Unknown function
SA0817	SA0817	hypothetical protein, similar to NADH-dependent flavin oxidoreductase	Cellular envelope and cellular processes
gmk	SA1052	guanylate kinase	Housekeeping - MLST
glpF	SA1140	glycerol kinase	Housekeeping - MLST
pbp2	SA1283	penicillin binding protein	Cellular envelope and cellular processes
aroE	SA1424	shikimate dehydrogenase	Housekeeping - MLST
aapA	SA1519	D-serine/D-alanine/glycine transporter	Cellular envelope and cellular processes
SA1544	SA1544	hypothetical protein similar to soluble hydrogenase 42kD subunit	Unknown function
SA1619	SA1619	hypothetical protein	ORPHANS
SA1621	SA1621	hypothetical protein	ORPHANS
hemH	SA1651	ferrochetalase homolg	Housekeeping
agrC	SA1843	accessory gene regulatorC	Information Pathways
leuB	SA1863	3- isopropylmalate dehydrogenase	Housekeeping
sigB	SA1869	sigma factor B	Information Pathways
luxS	SA1936	autoinducer 2 production protein luxS	Information Pathways
hutI	SA2121	imidazalonepropionase	Housekeeping
arcC	SA2425	carbamate kinase	Housekeeping - MLST
SA2439	SA2439	conserved hypothetical protein	Unknown function
SA2445	SA2445	hypothetical protein	ORPHANS
SArRNA16	SArRNA16	16S ribosomal RNA	Information Pathways

Table 2. Gene loci and categorisation. Genomic locations of selected loci are illustrated in Appendix A6.

3.2 RESULTS

3.2.1 Gene categories and functional constraint

The ratio of synonymous and nonsynonymous substitutions (d_S/d_N) has been calculated for each gene sequence as a measure of functional constraint and the strength of purifying selection within each gene.

Category	gene	d_S	d_N	d_S/d_N
CELLULAR ENVELOPE	*vicK*	0.0350	0.0000	α
	SA0272	0.0530	0.0070	7.6
	SA0817	0.0390	0.0060	6.5
	pbp2	0.0270	0.0030	9.0
	AapA	0.2660	0.0060	44.3
HOUSEKEEPING	SA0008	0.0110	0.0010	11.0
	SA0143	0.0120	0.0010	12.0
	SA0224	0.0290	0.0050	5.8
	yqiL	0.0310	0.0030	10.3
	pta	0.0250	0.0030	8.3
	tpi	0.0380	0.0060	6.3
	gmk	0.0350	0.0030	11.7
	glpF	0.0750	0.0050	15.0
	aroE	0.0410	0.0060	6.8
	hemH	0.0260	0.0020	13.0
	leuB	0.0260	0.0030	8.7
	hutI	0.0470	0.0020	23.5
	arcC	0.0190	0.0030	6.3
INFORMATION PATHWAYS	*sarA*	0.0010	0.0000	α
	serS	0.0130	0.0000	α
	dnaC	0.0500	0.0004	125.0
	SA0189	0.0230	0.0020	11.5
	tufA	0.0050	0.0010	5.0
	agrC	0.2020	0.0150	13.5
	sigB	0.0130	0.0010	13.0
	luxS	0.0290	0.0010	29.0
ORPHANS	SA0268	0.0140	0.0060	2.3
	SA0740	0.0240	0.0080	3.0
	SA1619	0.0760	0.0300	2.5
	SA1621	0.0650	0.0160	4.1
	SA2445	0.0450	0.0090	5.0
	SA0139	0.0380	0.0150	2.5
UNKNOWN FUNCTION	SA0778	0.0090	0.0000	α
	SA0013	0.0470	0.0000	α
	SA0100	0.0350	0.0000	α
	SA0775	0.0200	0.0000	α
	SA0275	0.0810	0.0030	27.0
	SA1544	0.0220	0.0030	7.3
	SA2439	0.0190	0.0180	1.1

Table 1. Frequencies of synonymous and nonsynonymous substitutions.

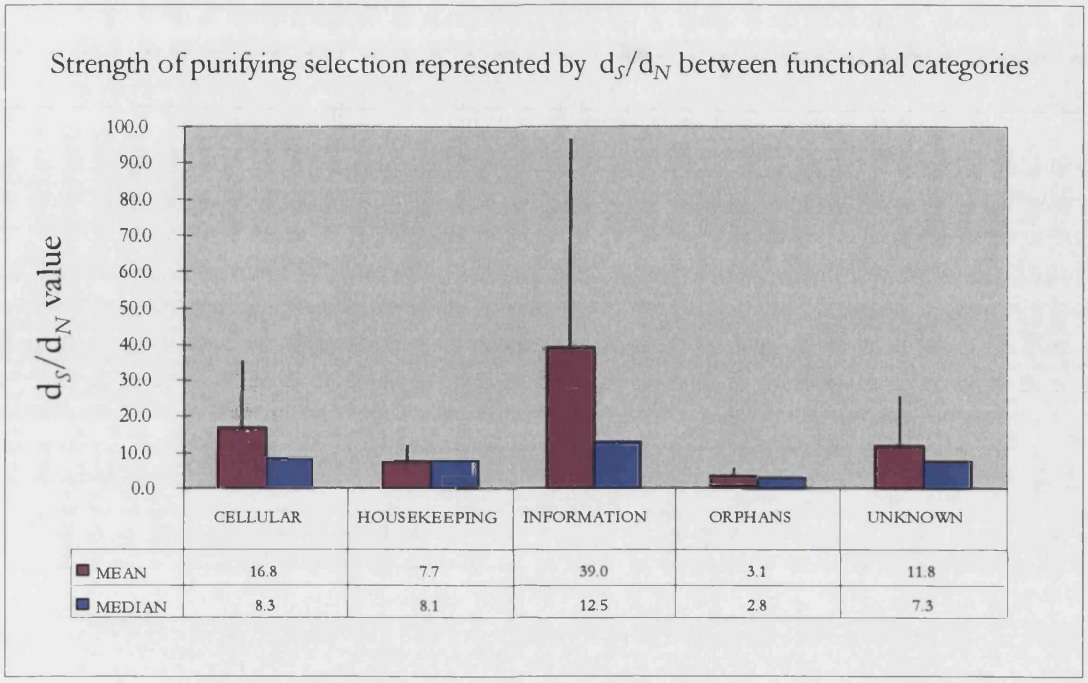


Figure 1. Mean and median d_S/d_N values for gene categories.

A higher value for the ratio of synonymous to nonsynonymous substitutions per site (d_S/d_N >1) means that there is an excess of synonymous substitutions and some proportion of the nonsynonymous substitutions have been removed by purifying selection. The mean and median values for the housekeeping genes and the ORPHANS are fairly consistent within these groups. This suggests that there is a fairly uniform level of purifying selection acting upon the genes included in each of these samples. The lowest mean d_S/d_N is for the ORPHAN genes and although the mean d_S/d_N >1 this suggests a much more relaxed functional constraint as purifying selection is much slower in removing nonsynonymous substitutions. The absence of these loci in closely related species suggests they are non-essential for bacterial cell survival, consistent with the low level of purifying selection observed here. We observe an excess of synonymous substitutions within the information pathway genes. However, the standard deviation (SD) indicates a high level of variation within this category although the median value is much smaller than the mean. The variability is in this category is largely attributable to a single gene, *dnaC*. In this case the extremely high d_S/d_N of 125 is a result of a large amount of synonymous variation with only two sites which are nonsynonymous. This is suggestive of extreme stabilising

selection consistent with the essential role of this gene in DNA replication. The cellular envelope and cellular processes genes are also more variable in d_S/d_N than the Housekeeping genes and ORPHANS where the values appear relatively consistent. The highest d_S/d_N of 44.3 is found for the gene *aapA* which is a D-serine/D-alanine/glycine transporter. The polymorphic sites for *aapA* are shown in Figure 2 and divide the strains into two distinct groups. The divergence between these two groups is extensive whereas strains in each of these two groups are quite uniform. This largely synonymous divergence accounts for the high d_S/d_N observed within this gene. Four genes of unknown function are unrepresented in the calculation of d_S/d_N due to an absence of non-synonymous substitutions. The variation found in d_S/d_N may or may not reflect the putative variation in function within this sample of genes since the function is unknown.

However, perhaps as a result of the variation of d_S/d_N within some categories, we find no significant difference in d_S/d_N values between different categories of genes using the Kruskal- Wallis test ($p>0.05$). Pairwise comparisons have also been conducted using a T-test to determine whether there is any significant difference between the d_S/d_N data for pairs of categories (Table 2). We find no significant difference in d_S/d_N for any pairwise category comparisons.

```
H417_st50     TAATCTACCC TCATCAGTAA TCCTCGTCAT CTACAGTTTA TTAATATGAG AGTAACACCT TGTGTT
H560_st121    .......... .......... .......... .......... .......... .......... ....C.
D274_st17     .......... .......... .......... .......... .......... .......... ....C.
D22_st182     .......... .......... .......... .......... .......... .......... ....C.
D456_st101    .......... .......... .......... .......... C......... .......... ....C.
D535_st59     .......... .......... .......... .......... .......... .......... ....C.
H402_st13     .......... .......... .......... .......... C......... .......... G...C.
D17_st20      .......... .......... .......... .......... C......... .......... G...C.
D547_st97     .......... .......... .......... T......... C......... .......... G...C.
emrsa3_st5    .......... .......... .......... T......... .......... ...G...... ....C.
emrsa4_st239  .......... .......... .......... .......... C......... .......... G...C.
H707_st49     .......... .......... .T....C... .......... .......... .......... ....C.
H116_st9      .......... .......... .......... .......... C......... .......... G...C.
H512_st1      .......... .......... .......... T......... C......... .......... G...C.
C2_st7        .......... .......... .......... .......... .......... .......... ....C.
H591_st8      .......... .......... .......... .......... C......... .......... G...C.
C437_st25     .......... .......... .......... .......... C......... .......... G...C.
D97_st55      .....C.... .......... .......... .......... .......... .A........ ....C.
H783_st15     .......... .......... .......... T......... .......... .......... ....C.
C101_st30     CTTCT.TTTT GTGATGACTC A.TATA.GGC .CGTGCAAAC .CG.CGATG. T.A.GTTTTC .TC.CG
H831_st36     CTTCT.TTTT GTGATGACTC A.TATA.GGC .CGTGCAAAC .CG.CGATG. T.A.GTTTTC .TC.CG
H19_st10      C..CT.TTTT GTGATGACTC A.TATA.GGC .CGTGCAAAC .CG.CGATG. T.A.GTGTTC .TC.CG
H295_st45     CTTCT.TTTT GTGATGACTC A.TATA.GG. .CGTGCAAAC .CG.CGATG. T.A.GTTTTC .TC.CG
C720_st22     CTTCT.TTTT GTGATGACTC A.TATA.G.C .CGTGCAAAC .CG.CGATGA T.A.TTTTTC .TC.CG
D470_st207    C..CT.TTTT GTGATGACTC A.TATA.GGC TCGTGCAAAT .CGGCGATT. G......A.. ...TCA
```

Figure 2. Variable sites within the 423 base sequence of *aapA*.

The dots in this represent site identity to the top reference sequence

	Cellular	Information	Housekeeping	ORPHANS
Cellular	-			
Information	p=0.527	-		
Housekeeping	p=0.467	p=0.082	-	
ORPHANS	p=0.09	p=0.137	p=0.06	-

Table 2. Significance of difference in d_S/d_N for pairwise category comparisons.

3.2.2 Gene categories and recombination

Population-scaled recombination rate (ϱ)

This is a measure of linkage disequilibrium. The likelihood of observing the data under a range of population-scaled recombination rates for each pair of segregating sites is estimated (Chapter 2, page 66). The population-scaled recombination rate (ϱ) with the highest likelihood is shown in Table 4. The highest value for ϱ, where the maximum value can be 100, is 34.343 for the information pathway gene SA0189. At least one value of zero is seen in all categories except the housekeeping genes and ORPHANS. An ANOVA has been used to test a null hypothesis of no association between gene category and the population-scaled recombination rate. The sum of squares = 46.9 with 4 degrees of freedom. This statistical test tells us that the probability of this result, assuming the null hypothesis, is 0.975. Therefore the differences observed in population-scaled recombination rate (ϱ) are not significantly different between the categorised genes in this dataset. Pairwise comparisons have also been conducted using a T-test to determine whether there is any significant difference between ϱ for pairs of categories (Table 3). We find no significant difference in ϱ for any pairwise category comparisons. This result tells us that the population recombination rate (ϱ) is not a predictor for gene category and that no gene from any category is more or less likely to have a higher population–scaled recombination rate than any other. The MLST genes and additional housekeeping genes have been grouped together for the purpose of this ANOVA but a T-test confirms that there is no significant difference in the data between these two sets of housekeeping genes with p=0.841.

	Cellular	Information	Housekeeping
Cellular	-		
Information	p=0.839	-	
Housekeeping	p=0.766	p=0.517	-
ORPHANS	p=0.881	p=0.689	p=0.899

Table 3. Significance of difference in population recombination rate (ϱ) for pairwise category comparisons

Minimum number of recombination events (R_M)

This parameter indicates the Minimum number of recombination events (R_M) in the history of the sample R_M is obtained using the four-gamete test (Chapter 2, page 65). The R_M has been calculated for each individual gene and the results are shown in table 5.

An ANOVA has been used to test a null hypothesis of no association between gene category and the Minimum number of recombination events (R_M). The sum of squares = 35.82 with 4 degrees of freedom. This statistical test tells us that the probability of this result, assuming the null hypothesis, is 0.349. Therefore the differences observed in the calculated Minimum number of recombination events (R_M) are not significantly different between the categorised genes in this dataset. The MLST genes and additional Housekeeping genes have been grouped together for the purpose of this ANOVA but a T-test confirms that there is no significant difference in the data between these two sets of housekeeping genes (p=0.632). A significant difference is observed in the data for housekeeping genes and ORPHANS. The mean for housekeeping genes is 2.38 whereas ORPHANS have a R_M mean of 5.00. There are a significantly higher Minimum number of recombination events in the ORPHANS than in the housekeeping genes included in this dataset.

	Cellular	Information	Housekeeping
Cellular	-		
Information	p=0.493	-	
Housekeeping	p=0.384	p=0.814	-
ORPHANS	p=0.267	p=0.119	**p=0.035**

Table 4. Significance of difference in Minimum number of recombination events (R_M) for pairwise category comparisons.

Sawyer's Runs Test, Bellerophon and Maximum Chi-squared

Mosaicism is the non random distribution of polymorphisms and is clearly illustrated in aligned variable sequences of the *hutI* gene in Figure 3. The presence of 25 variable sites within the last 168 bases of strain H466 compared to none in the first 639 bases is unlikely to have arisen by point mutation alone and would certainly be considered a result of recombination. However, not all recombination is as easily observed within sequences. Sawyer's Runs Test looks for evidence of recombinational exchange within aligned sequences by determining if regions of sequence pairs have more consecutive identical polymorphic sites in common than would be expected by chance. The SSCF is calculated from silent polymorphic sites only whereas the SSUF considers monomorphic sites (Chapter 2, page 63). The SSUF is more likely to detect larger gene conversions than the SSCF. The results for this test are shown in table 5. Bellerophon outputs the number of chimeric (mosaic) sequences and the identity of putative mosaic and parental sequences using a partial treeing approach. The significance of these mosaic and parental sequences can then be tested using the Maximum Chi-squared test (Chapter 2, page 64). The results of these tests are given in table 5.

```
          13344 55555 5555555555 66666 66666 6666666677 7777777777
          33634 12244 5666678889 00001 34556 6677999900 1134455679
          19090 32506 2145792581 03692 95140 6928036958 1721469412

D456      CGCGG AACTT AGGCATATCC AGGCT ACACC GTAACTTTCA CAGCGAAATT
H19       .AT.. ..... ..A.G..... ..... ..T.. .......... ..........
H402      G.... ..... .......... ..... ..... .......... ..........
H783      ..... ..... .......... ..... ..... .......... ..........
D97       .AT.. ..... .......... ..... ..... .......... ..........
D535      ..TAA ..... .......... ..... ..... .......... ..........
H466      ..... ..... .......... ..... GTTTT CCGTTACCAG TGCAATGGGC
c2        ..... ..... .......... ..... ..T.. .......... ..........
H591      ..... ..... .......... ..... ..T.. .......... ..........
D547      ..... GTTGC TA.A.ATATA GAATC GTTTT CCGTTACCAG TGCAATGGGC
H116      ..... ..... .......... ..... ..T.. ....T..... ..........
```

Figure 3. Mosaic sequence within the *hutI* gene.

The numbers read vertically as the site positions within an 807 base internal fragment of the housekeeping gene, *hutI*. Dots represent identity compared to the top reference sequence.

We observe evidence for recombination using Sawyer's Runs test in all gene categories in this data for both condensed (silent sites) and uncondensed (all sites) fragments. However, there appears to be relatively fewer genes of unknown function, and ORPHANS within the set of genes with significant results. Using Bellerophon there also appears to be a relatively even distribution of genes with evidence for recombination using these methods, although the information pathway genes have the fewest positive results.

Category	Gene	R_M	ϱ	Bellerophon	Max χ2	SSCF	SSUF
CELLULAR	*pbp2*	2	27.273	3	1	p> 0.05	
	SA0817	3	4.04	138*		p> 0.05	
	vicK	2	16.162				
	AtpA	4	0				
	SA0272	5	0	11	3	p> 0.001	p> 0.001
HOUSEKEEPING	*leuB*	4	4.04	1			
	hemH	7	21.212	4			
	SA0224	2	4.04	138*			
	hutI	2	5.051	138*	**	p> 0.001	
	SA0143	0	25.253				
	SA0008	1	3.03				
MLST	*aroE*	4	7.071				
	tpi	2	12.121				
	pta	1	5.051				
	yqiL	3	17.172	7			p> 0.05
	gmk	3	19.192			p> 0.05	p> 0.05
	glpF	1	2.02				
	arcC	1	17.172				
INFORMATION	*lucS*	0	4.04			p> 0.05	p> 0.05
	agrC	9	4.04				p> 0.05
	sigB	0	1.01				
	dnaC	4	17.172	8	2	p> 0.05	p> 0.05
	SA0189	3	34.343	2	2		
	serS	1	4.04				
	tufA	0	0				
	sarA	0	0				
ORPHAN	SA1619	11	27.273	7	2	p> 0.01	p> 0.01
	SA0740	6	4.04				
	SA2445	4	8.081	7			
	SA0139	3	6.061	6	1		
	SA0268	3	9.091				
	SA1621	3	8.081	7	1		
UNKNOWN	SA2439	3	8.081	138*			
	SA0775	0	3.03				
	SA1544	1	8.081	1	1		
	SA0100	2	11.111				p> 0.05
	SA0778	0	0				
	SA0275	11	12.121	2	2	p> 0.01	
	SA0013	6	32.323				

Table 5. Results for each test for recombination

* maximum default output for this program as no further details were given regarding recombinant and parental strains. ** indicates that manually observable mosaics are significant

3.2.3 Functional constraint and recombination

We have found no evidence for differential rates of recombination found within different gene categories, other than between housekeeping genes and ORPHANS by R_M. However, we have also found that there is no significant association in functional constraint, as represented by d_S/d_N , and gene category. Therefore d_S/d_N is not a predictor of gene category and vice versa. Despite the independence of category from recombination and d_S/d_N variables, there may be a relationship between recombination and functional constraint (d_S/d_N) , independent of gene category.

Regression analysis has been used to measure any relationship between functional constraint (represented by d_S/d_N) and population scaled recombination rate (ϱ) and the Minimum number of recombination events (R_M) (Figures 4 and 5). The calculated R^2 values are 0.0215 and 0.0053 respectively indicating that the differences in functional constraint only explain approximately 2% of the variation in recombination. In other words it appears that recombination is independent of functional constraint. The data for Sawyer's Runs test and Bellerophon cannot be displayed graphically. However, a T-test can be used to test the significance in difference of d_S/d_N between genes in which recombination was or wasn't detected using these methods. This test reveals that there is no significant difference in d_S/d_N between genes in which recombination has or has not been detected using these methods ($p= 0.061$ and 0.459 respectively). The relationship between functional constraint and recombination has been tested statistically for each method and in all cases the association between the two variables is not significant.

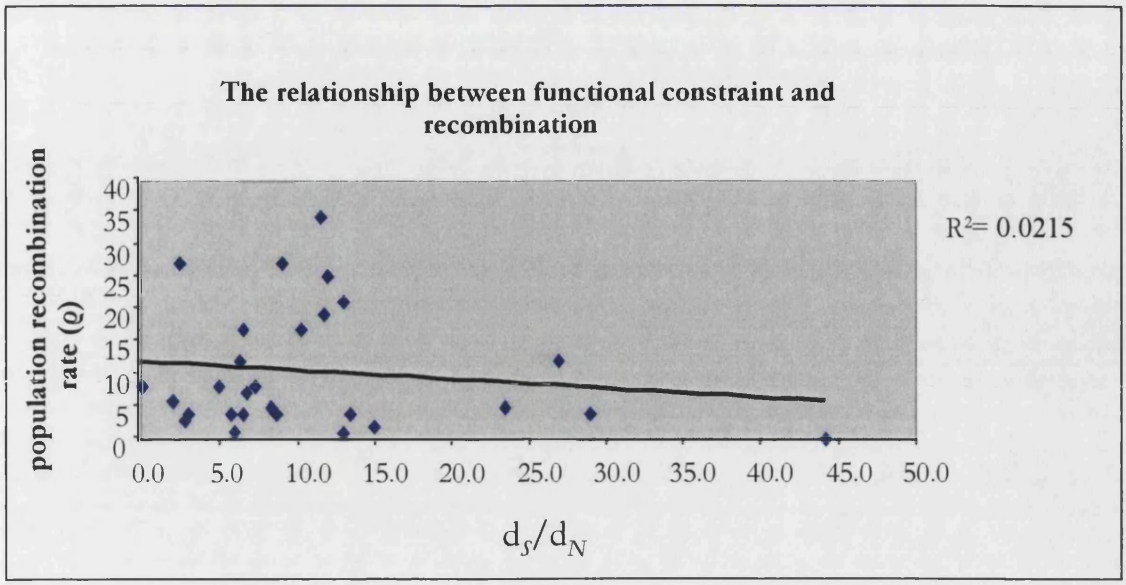


Figure 4. Population-scaled recombination rate (ϱ).

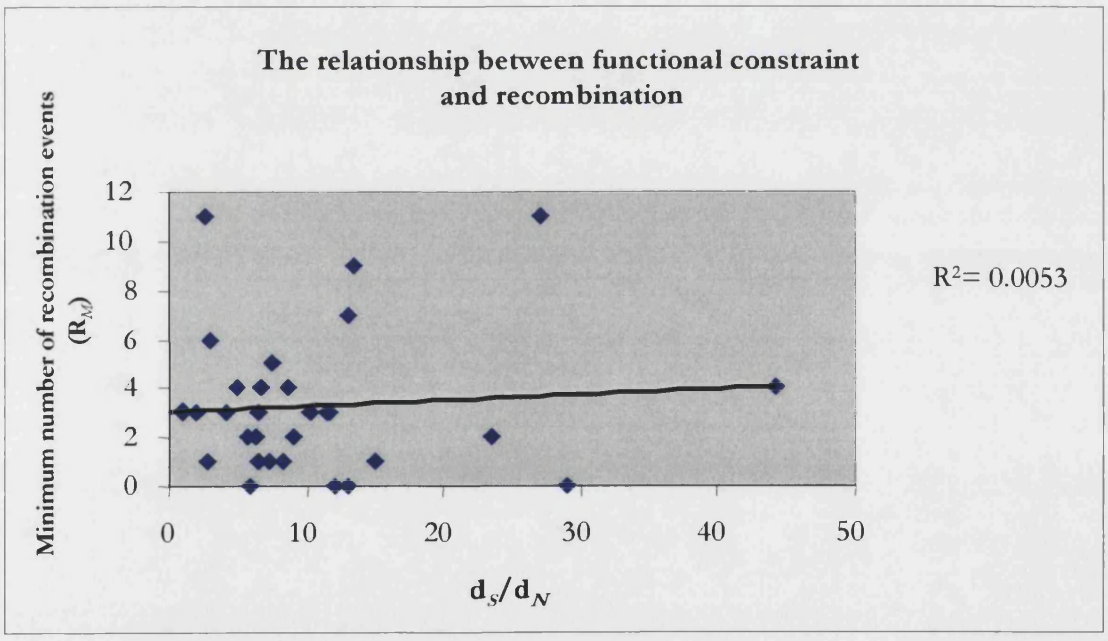


Figure 5. Minimum number of recombination events (R_M).

Measuring Recombination

An array of available tests have been used to establish the extent of recombination within this gene set. Not unexpectedly there are some conflicts in the outputs given for these tests since each examines the data in different ways (Table 5). For example, the regression between the Minimum number of recombination events (R_M) and the population-scaled recombination rate is only 0.097. The population-scaled recombination rate is calculated using coalescent likelihood of observing the data (for each pair of segregating sites) under a range of population recombination rates. The likelihoods are combined across pairs to provide a point estimate of the population recombination rate. Whereas R_M is calculated from a collection of correlated family trees generated for each site and naturally as a Minimum estimate will underestimate the true extent of recombination within the sample. The alternative tests Sawyer's Runs Test and Maximum Chi-squared examine the distribution of polymorphisms within the sequences to infer recombination. The universal problem with tests for recombination is sensitivity. Older recombination can be harder to accurately detect due to the subsequent diversification or amelioration of the original recombinant. Recombination between very closely related sequences can result in the substitution of very few sites in the recipient sequence. In such cases the distribution of polymorphisms can be a fairly insensitive measure and mosaics comprising just a few bases are less likely to be significant. Also for sequences with few informative sites a treeing approach can also be limited since the poor resolution makes comparisons difficult to accurately interpret. The genes *aapA* and *agrC* are both highly divergent (excess of polymorphisms within one or more groups of strains) loci with long branches within these individual gene trees. With the luxury of data from 37 other genes it is possible to identify divergent lineages within this population. From tree topology and visual inspection of the distribution of polymorphism we note that the length of internal branches within these loci is not typical. The distribution of polymorphisms across the entirety of the sequence is highly indicative of perhaps ancient recombination within these genes. If we had sequenced further around these regions we may find a cut-off point for such a putative event whereby the level of divergence returns to that predicted by all other data. However, there is little suggestion for this from the data collected. The significance of the uncondensed fragment (SSUF) which accounts for all polymorphic sites (not just segregating sites) using Sawyer's Runs Test for *agrC* however is suggestive of a larger

recombination event. If we look at all the data we find that in the absence of information from other loci this is completely undetected by all tests for *aapA*. For *agrC*, this is overlooked by Bellerophon. The estimates for Bellerophon are likely to be underestimates of the true extent of recombination within the submitted sequences. This program only outputs mosaics where two putative donors are present within the sequences. In this way unique mosaics are overlooked. For example, the unique mosaic in strain H295_st45 of gene SA0817 (Figure 6) is not included in the output since no putative donor for that sequence can be identified within the sequences. The smallest window size for this partial treeing approach implemented in Bellerophon is 200 bases. Where sequences are only 400-500 bases in length the implementation of the method uses only 2 trees thus limiting the power of the approach. The Maximum chi-squared test in itself is also limited by the fact that no more than 3 sequences (1 recombinant and 2 parental) can be submitted at one time. In this way it is limited for population scaled studies. However, teamed with Bellerophon which identifies both the putative and recombinant sequences this limitation is overcome and it can be used to its full potential to determine the significance of the mosaics.

```
C101_st30      CATCATAAGG ACGTAGGGTC AAACAGTGTG TACACAAGCC AGGGACCCGA
H831_st36      .......... .......... .......... .......... ..........
D22_st182      ..A....... ........AT ..GA..G.C. C......... ...A......
D274_st17      .......... .....A..A. ..G...G.C. CG........ ...A......
D535_st59      .......... .....A..A. ..G...G.C. CG........ ...A......
D365_st121     .......... ........A. ..G...G.C. C........T ...A.T....
C2_st7         ...........          TAA. ..GT..G... C......... ..TA......
H295_st45      ...GAGCCA GAAAG..A. ......GA.. .......... ...AT.....
C437_st25      .......... .....A..A. ..GT..G... C..TTGGA.. ...A......
H116_st9       ...T...... ........A. ..G..AG.C. C.A....... ...A.....G
emrsa4_st239   ...T...... ........A. ..G..AG.C. C......... ...A......
H591_st8       ...T...... ........A. ..G..AG.C. C......... ...A......
H783_st15      ...T...... ........A. ..G..AG.C. C......... ...A......
C640_st22      .......... ........AT T.G...G.C. C....GGA.. ...A......
D17_st20       .......... .....A..A. ..G...G.C. C......... G..A......
D470_st207     T......... ........A. ..G...G.C. C...TGGA.. ...A.....G
H417_st50      .......... ........A. ..G...G.CA C......... .A.A.T....
H707_st49      .......... ........A. ..G...G.CA C......... ...A...TA.
D97_st55       .G........ ........A. ..G...G.C. C......... ...A.TA...
H466_st5       .......... .....A..A. ..G...G.C. C......... ...A....A.
H19_st10       .......... ........A. ..G...G.C. C......... ...A......
D456_st101     .......... .......AA. .TG.G.G.C. C......... G..A......
H512_st1       .......... .......AA. .TG.G.G.C. C......... G..A......
D547_st97      .......... .......AA. .TG.G.G.C. C.......T. G..A......
H402_st13      .......... .....A.TA. ..G...G.C. C......... G..A......
```

Figure 6. Variable sites within SA0817 sequences.

3.2.4 *Staphylococcus aureus* phylogeny

Having used various tests for recombination, we find fewer genes for which there is evidence for recombination, than those where there is no evidence of recombination. The population-scaled recombination rate values shown in table 1 range from 0 - 34.343 where the maximum value can be 100. This intermediate level of recombination suggests that the phylogenetic signal will be sufficiently maintained to feasibly generate a robust phylogeny for *S. aureus*. Thirty strains representing the diversity of the natural population of *S. aureus* are included in a Bayesian reconstruction (Chapter 2, page 61) shown in Figure 7. This consensus phylogeny represents the concatenated sequences of 37 genes representing 17.8 Kb of nucleotide sequence. The increase in informative sites provides a much higher level of resolution between strains and clonal complexes previously identified in Feil *et al.*, 2003. Founder STs of these clonal complexes are shown in Figure 7 by shaded circles. The posterior probabilities provide strong support for most of the branches in the this tree, the lowest support is found within the lower portion of the tree where despite the large amount of sequence data is still less well resolved. The multifurcating region shown by the Bayesian tree generated for the concatenated sequence of 7 MLST loci has been completely resolved. This resolution supports the major division of *S. aureus* previously identified and further resolves a third clade. These population dividing branches are supported by posterior probabilities of 100 and are shown in red. The same data is also used in a Splits decomposition analysis in which alternative phylogenies will also be represented by parallel lines (Figure 8). This Splits tree supports both the low level of recombination in this species, a good level of phylogenetic agreement, branch lengths determined by Bayesian analysis and the division of the population into 3 groups. The population framework observed is suggestive of 3 successful ancestral *S. aureus* lineages which have subsequently diverged to form the structure we observe in Figures 7 and 8. The data for all genes of all categories was subdivided into the three groups, as shown in Figure 7, for further analysis regarding the evolutionary significance of such a division. The d_S/d_N and population-scaled recombination rate (ϱ) then was calculated for each of the 3 subsets of data for each gene (data in appendix A2-4). An ANOVA was then used to test the significance of all values for all genes between the three groups. This test showed that there is no significant difference in d_S/d_N between the three groups in this dataset.

($p=0.058$). There was also no significant difference in population-scaled recombination rate between the three groups using an ANOVA ($p= 0.99$). The difference in branch lengths in Group 1 compared to Group 2 is quite striking. ST36, for which a complete genome has been sequenced, falls within group 1 and has been reported as a much more divergent genotype (Holden *et al.*, 2004). In contrast Group 2 branches are much shorter and relationships are less clear. This observation could be a result of one of two things: recombination within this group, or recent divergence. The lack of evidence for more recombination within group 2 suggests that this may reflect more recent divergence. Individual gene topologies for the Group 1 lineages confirm that ST36, ST45, ST10 and ST207 are quite uniquely divergent except for ST22 which can be observed sharing branches with unrelated STs suggesting a history of recombination in this genotype.

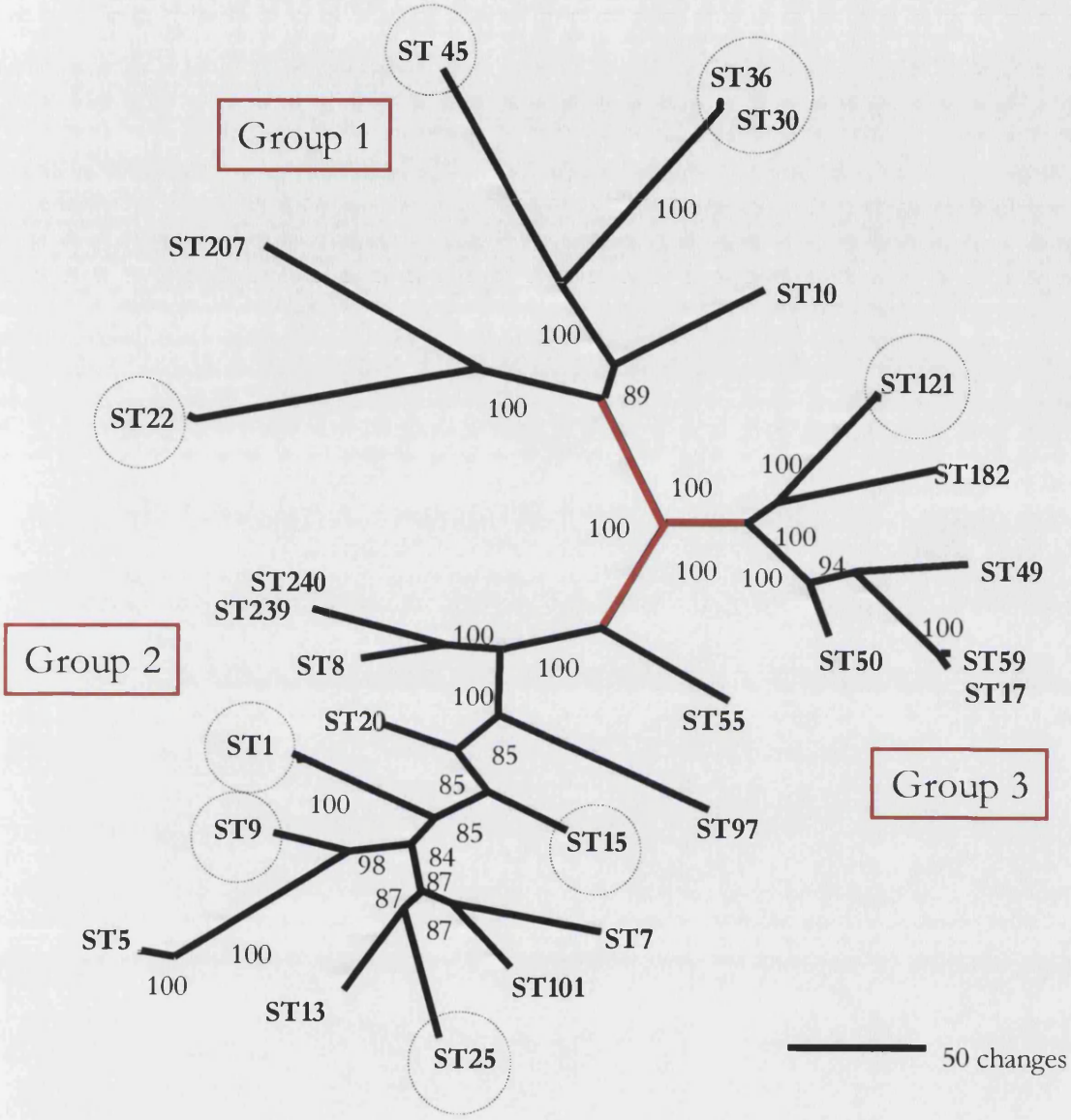


Figure 7. Bayesian phylogeny for *Staphylococcus aureus*.

The location of clonal complexes in the context of the population framework is shown on the tree by grey dashed rings.

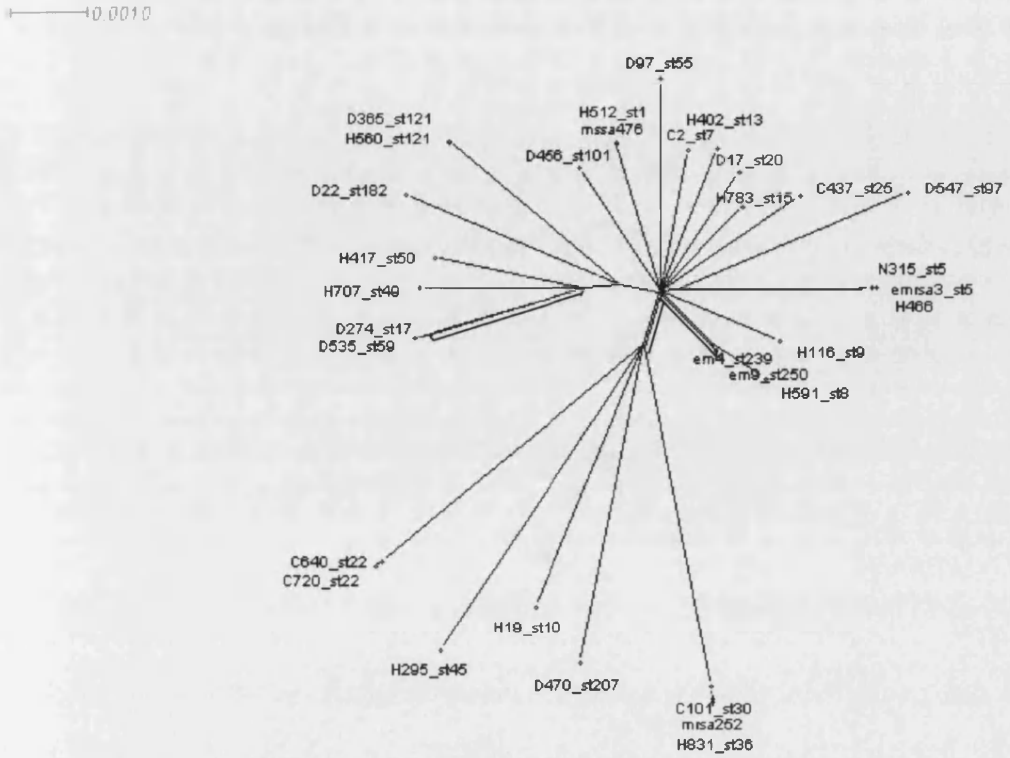


Figure 8. Consensus Splits Decomposition for *S. aureus*.

Parallel lines are observable between lineages ST239 and ST30 and between ST59 and ST17. Robinson and Enright report the existence of two large replacements of ~244 and 557 kb within *S. aureus* lineages. A recombinant lineage, ST239, and parental lineages ST8 (Group 2) and ST30 (Group 1) are included in this study. ST239 is an SLV of ST8. The recombination replacement spans SA2339, includes the origin of replication and ends within SA0318. Within this region ST239 has almost complete identity to ST30. However, in remainder of the genome this lineage resembles its ancestor by descent ST8 (Robinson & Enright, 2004). The relationships between these lineages within individual gene trees change according to the genomic location of the gene, consistent with the existence of this replacement. The parallel lines in this Splits Tree reflect this novel evolutionary event. STs 59 and ST17 are very closely related and the lack of discriminating sites between these two STs presumably compromises confidence in the assignment of this branch in the context of longer evolutionary relationships.

Orthologous sequences for the *S. aureus* genes sequenced in this chapter were identified in *S. epidermidis* ATCC 12228 (Zhang *et al.*, 2003) where possible to be used as outgroups. In the light of the three groups identified within the *S. aureus* population such outgroup sequences could help determine whether an ancestral group exists. However, a high level of identity and conservation between *S. aureus* and *S. epidermis* sequences is also reflected in a high level of conservation and small internal branches within *S. aureus* sequences. This means that the resolution within the *S. aureus* sequences is insufficient and the integrity of the groups is lost with the inclusion of *S. epidermidis* (Figure 9b). A fine balance of intraspecifc divergence and interspecific conservation is required. In the cases where there is atypical divergence and longer internal intraspecific branches (Figure 9a) these branches are maintained even with the inclusion of *S. epidermidis* sequence as an outgroup. In this case the *S. epidermidis* branch appears to lie fairly central to group 1 and groups 2 and 3 and does not root the intraspecific sequences by any group over another.

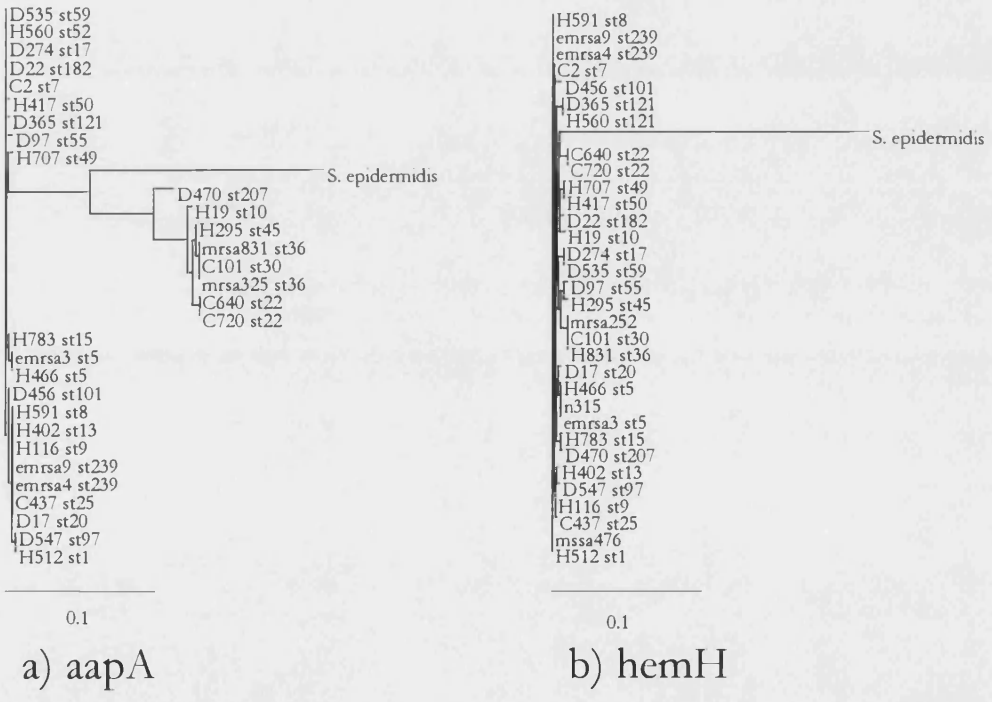


Figure 9. Gene trees rooted with *S. epidermidis* orthologous sequence.

3.2.5 Phylogenetic congruence and reliability

Congruence refers to the consistency of phylogenetic signal between unlinked loci. Congruence is used here to score the topologies of individual gene loci against the concatenated data and compare these to the 99th percentile of scores obtained for trees of random topology. Genes which have a higher log ML score than the 99th percentile score of random trees are deemed congruent. These scores are shown in Table 9. The first tree listed in the table is the tree generated from the concatenated data, scored against the concatenated data. As would be expected we find the highest likelihood, represented by the lowest log likelihood score, for this tree. There are only three genes which are scored as no more congruent with the concatenated tree as trees of random topology (Table 9). The topologies of these three genes are shown below in Figure 10.

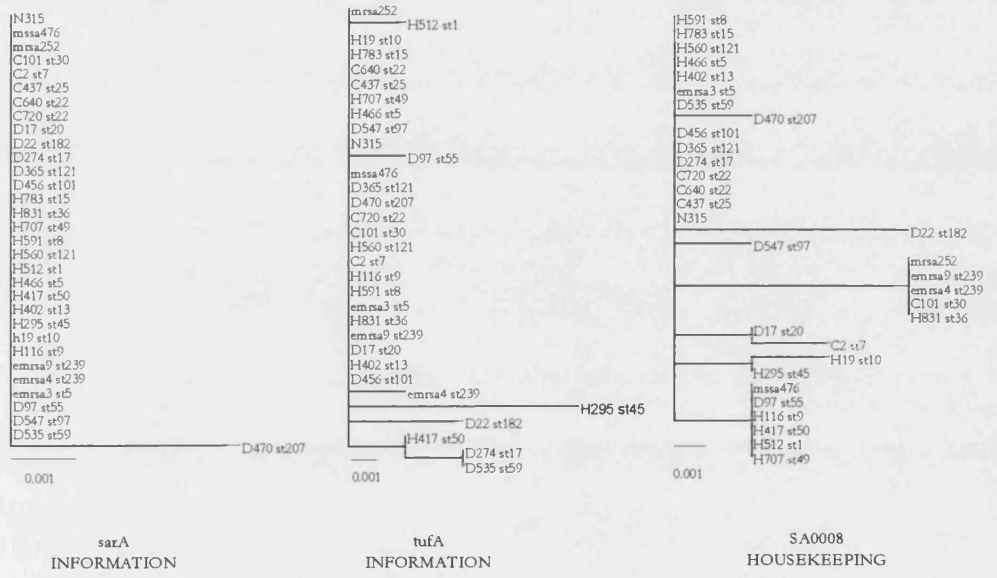


Figure 10. Topologies of incongruent loci.

The topologies of these trees are highly uninformative regarding relationships between strains. This is a reflection of the paucity of informative sites found within these sequences. For example within the information pathway gene *sarA* we find only 2 alleles which differ by only one site in one strain. This incongruence is not a result of extensive recombination as no recombination is detected in *tufA* or *sarA* and a very low measure in SA0008 by population scaled recombination rate (ϱ) and Minimum number of

recombination rate (R_M) (Table 5). Maximum likelihood score presents us with a ranking of phylogenetic reliability between genes (Table 7). However, this is not an independent ranking of genes since the data for each of these genes is in the data they are scored against. A gene with a higher proportion of informative sites will have more weight in the data than a gene with few variable sites. Instead we can use the SH test score, which represents the difference between ML scores for the test data and the highest scoring tree. The test data is removed from the consensus data thus providing an independent scoring of phylogenetic reliability (Chapter 2, page 67). The gene sequence with the best independent fit to the tree of the remaining data is SA2439 (Table 8). This is a gene of unknown function. The best fitting three genes represent genes of unknown function, cellular envelope and cellular processes and the ORPHANS. A housekeeping gene is the fourth best fitting gene. The highest scoring of the information pathway genes is in place 10. An ANOVA tells us that there is no significant difference in SH scores ($p=0.078$) between the categories. The MLST genes and additional housekeeping genes have been grouped together for the purpose of this ANOVA but a T-test confirms that there is no significant difference in the data between these two sets of housekeeping genes with $p=0.548$. Pairwise comparisons between categories using the T-test also reveal that there is no significant difference in SH scores between pairwise category comparisons.

	Cellular	Information	Housekeeping
Cellular	-		
Information	p=0.058	-	
Housekeeping	p=0.095	p=0.292	-
ORPHANS	p=0.712	p=0.062	p=0.127

Table 6. Significance of difference in SH scores for pairwise category comparisons.

Tree	Log ML score	Category
CONCATENATED	36617.5	ALL CATEGORIES
SA2439	37890.4	UNKNOWN
pbp2	38153.2	CELLULAR
SA1619	38192.9	ORPHAN
SA0740	38356.3	ORPHAN
leuB	38371.4	HOUSEKEEPING
SA0775	38646.4	UNKNOWN
hemH	38706.4	HOUSEKEEPING
SA1544	38856.9	UNKNOWN
SA0224	38896.4	HOUSEKEEPING
agrC	39080.2	INFORMATION
luxS	39101.9	INFORMATION
SA0817	39121.6	CELLULAR
SA2445	39126.5	ORPHAN
hutI	39144.5	HOUSEKEEPING
vicK	39161.9	CELLULAR
aapA	39234.9	CELLULAR
aroE	39246.1	MLST-house
tpi	39614.5	MLST-house
sigB	39620.3	INFORMATION
dnaC	39631.6	INFORMATION
SA0100	39648.7	UNKNOWN
pta	39743.6	MLST-house
SA0139	39815.2	ORPHAN
SA0275	39870.3	UNKNOWN
SA0268	39895.7	ORPHAN
SA0778	39970.6	UNKNOWN
SA0013	40011.2	UNKNOWN
yqiL	40020.7	MLST-house
SA0272	40248.7	CELLULAR
SA0189	40323.6	INFORMATION
gmk	40605.9	MLST-house
glpF	40899.8	MLST-house
SA1621	40939.8	ORPHAN
aroC	40993.8	MLST-house
serS	41251.2	INFORMATION
SA0143	41337.8	HOUSEKEEPING
SA0008	**42081.9**	HOUSEKEEPING
tufA	**42350**	INFORMATION
sarA	**42997.5**	INFORMATION

Table 7. Congruence analysis for individual gene trees against the concatenated data.

Rank	Gene	SH score	Category
1	SA2439	1345.368	UNKNOWN
2	*pbp2*	1565.546	CELLULAR
3	SA1619	1708.409	ORPHAN
4	*leuB*	1776.105	HOUSEKEEPING
5	SA0740	1800.583	ORPHAN
6	SA0775	2056.379	UNKNOWN
7	*hemH*	2116.149	HOUSEKEEPING
8	SA1544	2255.252	UNKNOWN
9	SA0224	2340.697	HOUSEKEEPING
10	*luxS*	2507.735	INFORMATION
11	SA0817	2536.186	CELLULAR
12	SA2445	2573.84	ORPHAN
13	*vicK*	2604.608	CELLULAR
14	*hutI*	2627.56	HOUSEKEEPING
15	*aapA*	2671.307	CELLULAR
16	*aroE*	2677.064	MLST- house
17	*agrC*	2838.84	INFORMATION
18	*sigB*	3011.601	INFORMATION
19	*tpi*	3043.117	MLST- house
20	*dnaC*	3080.7	INFORMATION
21	SA0100	3109.434	UNKNOWN
22	*pta*	3151.953	MLST- house
23	SA0139	3220.862	ORPHAN
24	SA0268	3271.318	ORPHAN
25	SA0778	3364.755	UNKNOWN
26	SA0275	3388.939	UNKNOWN
27	*yqiL*	3436.345	MLST- house
28	SA0013	3458.439	UNKNOWN
29	SA0189	3757.817	INFORMATION
30	*gmk*	4010.359	MLST- house
31	*glpF*	4307.895	MLST- house
32	*arcC*	4459.983	MLST- house
33	*serS*	4661.175	INFORMATION
34	SA0143	4767.796	HOUSEKEEPING
35	SA0008	5480.02	HOUSEKEEPING
36	*tufA*	5727.689	INFORMATION
37	*sarA*	6564.651	INFORMATION

Table 8. Ranking congruence of single gene loci with the concatenated data using the SH test.

The log ML scores for individual category trees (generated from concatenated data of all genes representing that category) against the data of another category (concatenated data of all genes representing that category) is shown in Table 9. If the log likelihood score for a given topology is greater than that of the 99th percentile of scores for 200 trees of random topology then that topology is deemed no more congruent to the data than trees of random topology. In all cases, the log likelihood score for the concatenated category trees is lower than that for the 99th percentile of the distribution of 200 random trees. In other words, trees for different gene categories are significantly more similar to each other than to trees of random topology.

INFORMATION data	
Tree	log ML score
UNKNOWN FUNCTION	5958.79948
NO SIMILARITY	5997.947
HOUSEKEEPING	6002.06561
CELLULAR	6024.27317
random trees 99th percentile	6569.67792
NO SIMILARITY data	
Tree	log ML score
HOUSEKEEPING	6368.57202
UNKNOWN FUNCTION	6391.25944
CELLULAR	6424.19367
INFORMATION	6505.50123
random trees 99th percentile	7592.1237
HOUSEKEEPING data	
Tree	log ML score
NO SIMILARITY	13073.314
CELLULAR	13076.7723
UNKNOWN FUNCTION	13092.33066
INFORMATION	13136.14054
random trees 99th percentile	14191.3534
CELLULAR data	
Tree	log ML score
HOUSEKEEPING	4914.03809
UNKNOWN FUNCTION	4993.21429
NO SIMILARITY	5041.26024
INFORMATION	5078.5434
random trees 99th percentile	5833.2648
UNKNOWN FUNCTION	
Tree	log ML score
NO SIMILARITY	6166.03117
HOUSEKEEPING	6179.35628
CELLULAR	6193.41908
INFORMATION	6206.25206
random trees 99th percentile	6759.05856

Table 9. Congruence analysis for concatenated category trees.

We can also rank functional categories in the same way using the SH test. For example, to obtain an independent likelihood score for a tree generated from the concatenated information pathway genes alone, we score this against the data and tree for all the data except the information pathway genes. The category tree which best fits the data from all other categories is the tree for genes of unknown function. The tree for the cellular envelope genes, and the housekeeping genes also score highly. The lowest scoring category is represented by the MLST housekeeping genes tree (Figure 11). Additional housekeeping genes and MLST housekeeping genes are separated here to represent a typically sized dataset for an MLST typing scheme. It is interesting however that there is such a striking difference between the two sets of data which represent genes from the same category. One possibility is that these differences reflect a greater number of informative sites in the non-MLST housekeeping genes as ~800 bp was sequenced in 3 of these genes. However, the overall concatenated sequences do not differ greatly (3.2 Kb for MLST genes, 3.8 Kb for non-MLST genes) and the number of informative sites is also similar which means this is unlikely to account for the observed difference in phylogenetic reliability.

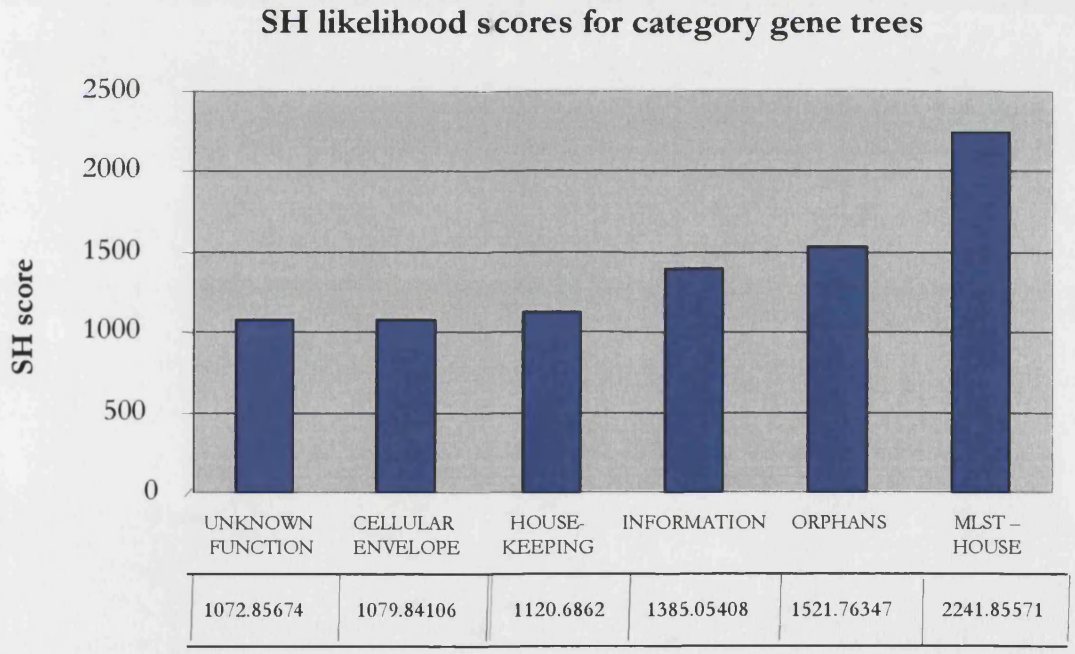


Figure 11. Graphical representation of the ranking concatenated category loci against all other categories using the SH test.

3.2.6 Predicting phylogenetic consistency

Although the weighting of a gene within a concatenated tree has been controlled for by its removal from the data set and tree it is scored against (SH test). The number of informative sites provided by a gene sequence will invariably affect its ability to reflect the phylogeny of all other genes. Figure 12 shows the relationship between the SH score for each gene and its pairwise diversity. The top three genes SA2439, *pbp2* and SA1619 are very different in their average pairwise diversity. These are indicated by dashed circles, as are the two most divergent loci, *aapA* and *agrC*. However, there does appear to be a cut-off which is indicated by the dashed line. Beyond this point is lie loci with less than 3% variation. This suggests that the decrease in SH score for these loci may be attributable to poor resolution as a result of their lack of variation. The regression value for the slope is 0.137. This means that only 13.7% of the result for SH score can be explained by average pairwise diversity. There are clear outliers in the trend represented by *aapA* and *agrC*. These have a high average pairwise diversity and yet are ranked 15th and 17th respectively. Thus, there is a trade-off: too much variation splits branches and not enough variation does not split branches which should be split.

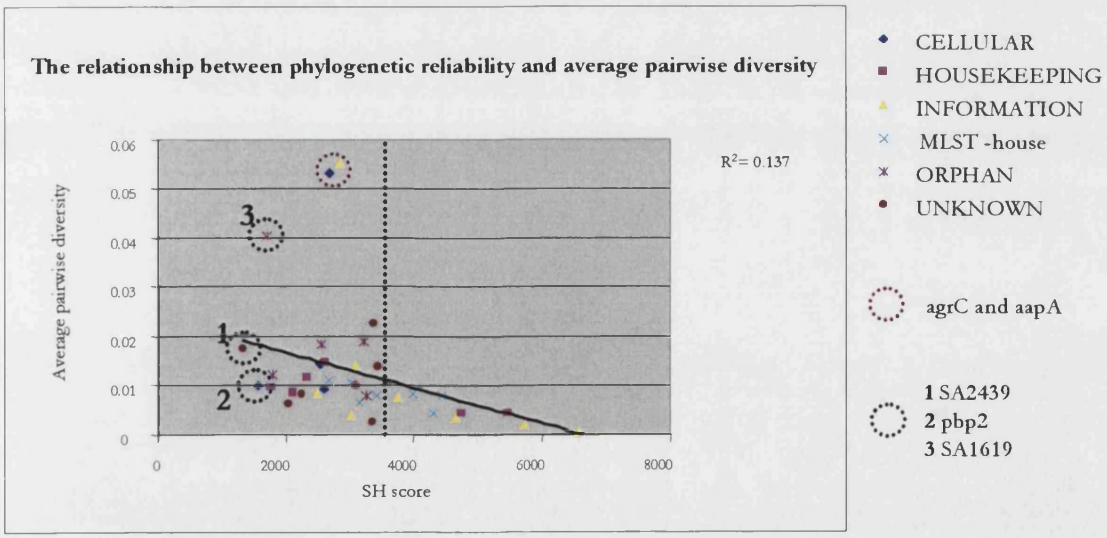


Figure 12. The relationship between average pairwise diversity and phylogenetic consistency.

We can test the association between phylogenetic reliability and recombination using regression. Figure 13 shows a scatter plot for SH score and population scaled recombination rates (ϱ). We observe a high amount of scatter within this plot and a regression value of only 0.0119. The calculated population-scaled recombination rate (ϱ) appears to account for less than 2% of the variation in the phylogenetic consistency of loci. This tells us that there is no relationship between phylogenetic reliability and population-scaled recombination rate (ϱ). We also find no significant association between the SH score and the Minimum number of recombination events (R_M). Again, there is a high amount of scatter within the scatter plot in Figure 14 and a regression value of 0.1459. This tells us that only 14% of the variation in the phylogenetic consistency of loci can be explained by the Minimum number of recombination events (R_M). In other words there is no relationship. In both these plots we observe a tail off whereby beyond a certain point the values with the lowest scores by the SH test have a low measure of recombination. These are, in fact, the very uniform and uninformative genes where the tests for recombination become less sensitive.

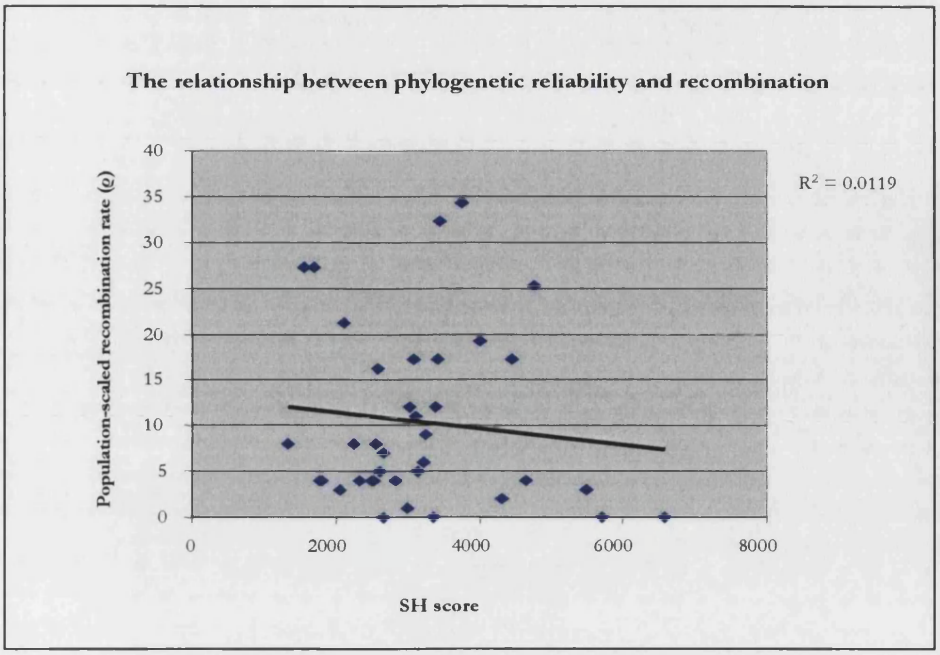


Figure 13. Association between population-scaled recombination rate (ϱ) and phylogenetic consistency of single loci (as estimated using the SH test).

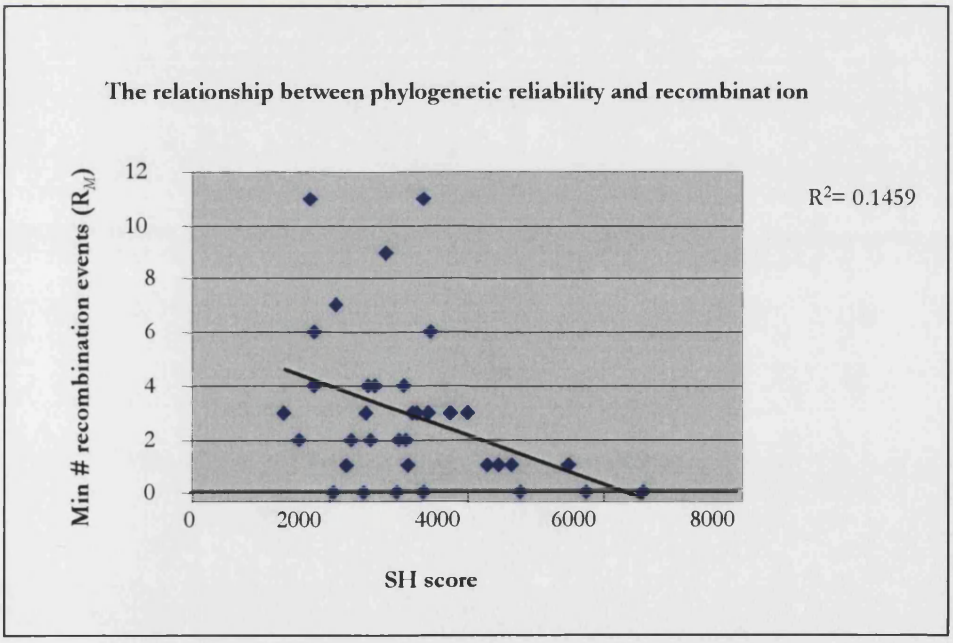


Figure 14. Association between Minimum number of recombination events (R_M) and phylogenetic consistency of single loci (as estimated using the SH test).

3.2.7 Genomic location

There are various other factors regarding gene location which may influence both the phylogenetic reliability. These may include proximity to the origin of replication and coding strand. In Figure 15 we test the association between phylogenetic reliability and distance from the origin of replication. There is a great deal of scatter of this plot and the regression tells us that distance from the origin accounts for <6% of the variation in SH score. The removal of the three lowest scoring genes which were found to be more congruent with the consensus tree than trees of random topology adjusts this regression value to only 0.0715. There is also no significant difference between the SH scores for loci on the leading and lagging replication strands p= 0.843.

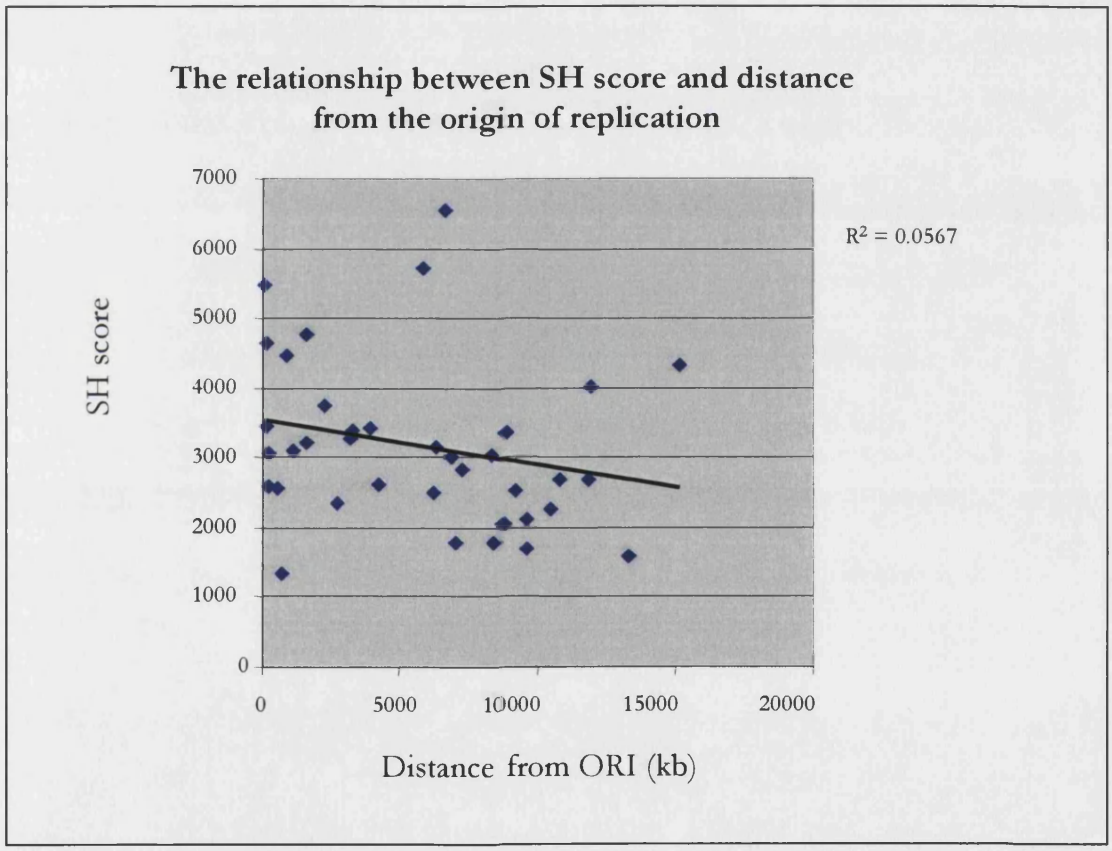


Figure 15. The relationship between SH score and distance from the origin of replication (ORI).

So far the results have demonstrated that differing gene loci have experienced a range of recombination rates, exhibit differing levels of diversity and are subject to varying selective constraint although there is no obvious explanation for these differences in terms of gene function or distance from the origin. It is possible that stochastic effects play a major role in determining the patterns of sequence evolution within specific loci. It is therefore pertinent to examine how localised these effects may be and whether clusters of linked genes display similar patterns of evolution and phylogeny. The current dataset contains 2 sets of closely linked genes. Firstly, SA0268, SA0272 and SA0275 are found within an 8.5 kb range. Relatively low levels of recombination are found within these three genes and yet the tree topologies for these gene nucleotide sequences (Figure 16) show differences in relationships between lineages. This unlikely to be solely due to a lack of informative sites as the % diversity in these genes is reasonably high (0.7%, 2% and 2%). This suggests that this recombination has occurred between small, localised fragments.

Secondly, genes SA1619 and SA1621 are found in a smaller 1.7 kb range. There is a fairly consistent low level of purifying selection acting upon these two ORPHANS (d_S/d_N - 2.53 and 4.063 respectively). However, yet again there are clear topological differences between trees for these two genes (Figure 17). Evidence for recombination within each locus has been seen in Table 5 and presumably accounts for some of the localised differences observed here between trees. We also observe both small and large mosaics between unrelated lineages (Figure 18a). However, this evidence for recombination also includes the presence of a significant mosaic found within both these genes. There is 100% homology between strains D17_st20 and H591_st8 in the terminal part of the SA1619 sequence and continuing into SA1621. The most parsimonious explanation would be that a single recombination event estimated at approximately 1.041 kb has occurred between these strains encompassing both loci SA1619 and SA1621, and the gene located between. These events are more clearly represented in Figure 18b.

These results show some localised differences in sequence evolution, tree topology and recombination between closely linked genes. We find evidence for only one larger-scale recombination event within genes SA1619 and SA1621. It appears likely that the majority of recombination events are small and localised and that larger events are much rarer.

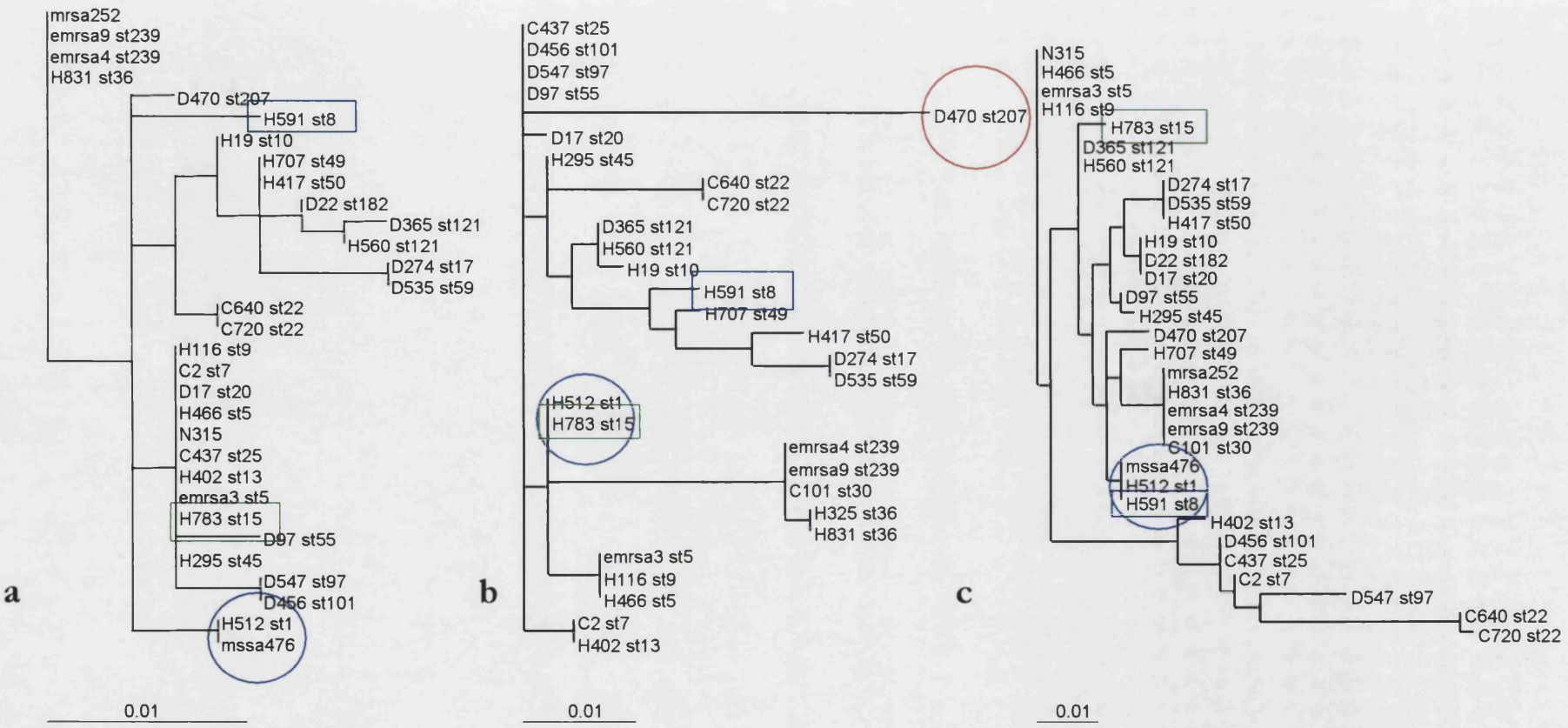


Figure 16. Maximum likelihood topologies for neighbouring loci a) SA0268 (ORPHAN) b) SA0272 (Cellular) and c) SA0275 (unknown)

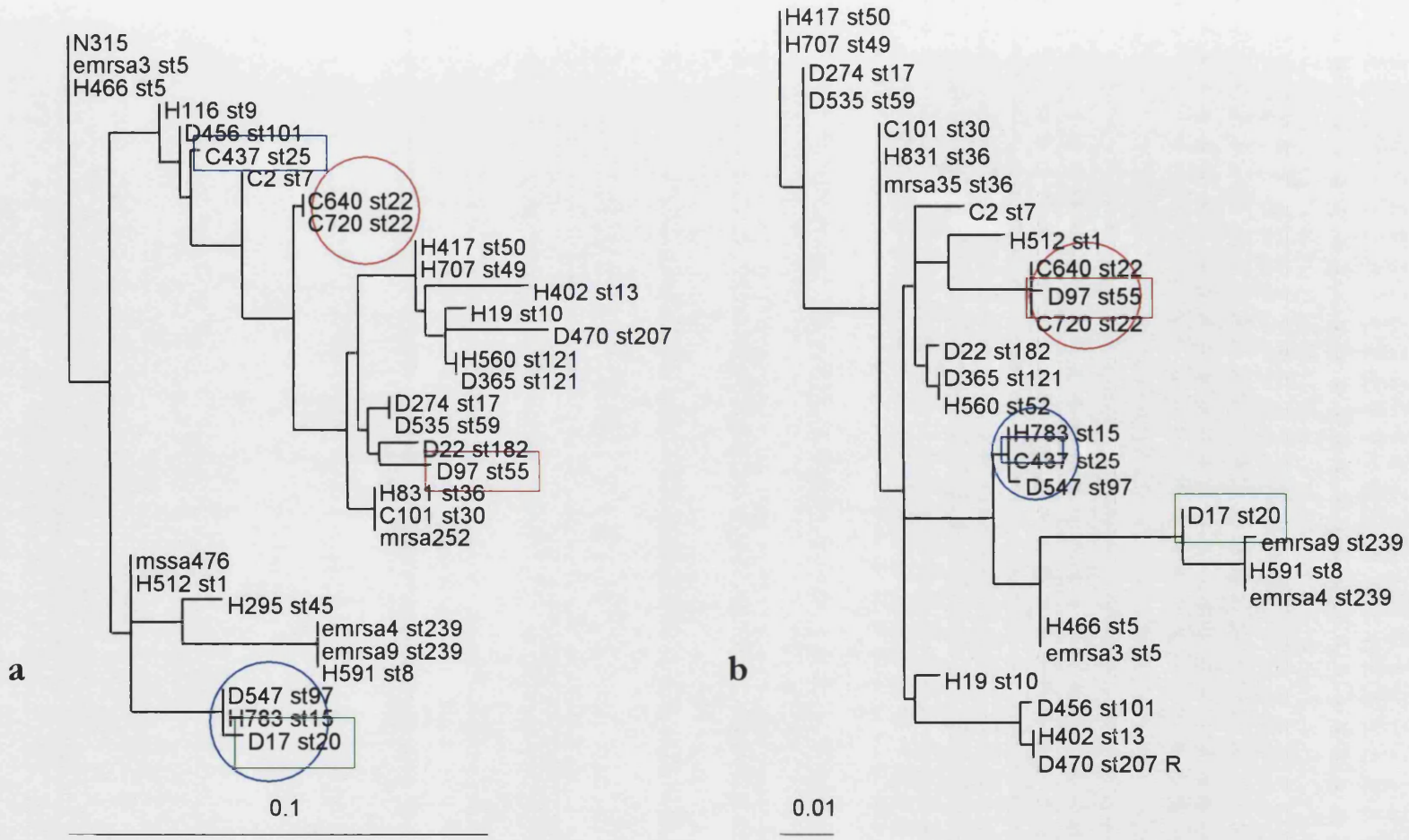


Figure 17. Maximum likelihood topologies for neighbouring loci a) SA1619 (ORPHAN) b) SA1621 (ORPHAN)

a

```
                    1111111 1111111122 2222222222 2333333333 3333333444 44 |      111 1111111111 2222222222 3333333333 3334444444
             112234445 6680112223 3334577800 0023445667 8034566667 7777889011 11 |1234679234 4566778899 0233444677 1122333567 8891233445
            2680375895 5912273591 5897547412 5751369790 1698123491 3789230313 47 |5774983493 8025170356 4078012404 8914356745 5658069156
C720_st22   AACCACCCAA CCACGCAGTC CAAAAACCAC CACATTCTCG TTCCCGTGAC GTGGAAGCTT TC|TTCACGAAAG ACTTCCACCT CGCACTATCT TGTTCCGCAT AGAGCTTGAA
H512_st1    .G.......C .......... ......T... ...TC..... ........TA .CA.GGA.CA .T| AC....G... GT........ ..........CA.CTT.... ..........
D17_st20    .G...G...C .A..ATT.CT T.T....... ...TC...A. ........TA .CA.GGA.CA .T|A..GTAGT.. ...CT.G.TC TA.G.GTGTC CA.CTT.... ..C.......
emrsa3_st5  .G..CG...C .A........ ...T.GT... ...TC..... ........TA ACA.GGA.CA .T|......G... ......G... .......... CA.CTT..G. G..T....G.
H591_st8    GG.......C .A.A....CT T.T....... .......C.A .C..TACATA .CA.GGA.C. .T|A..GTAGT.. ...CT.G.TC TA.G.GTGTC CC.CTTA... G.......G.
H783_st15   .G.......C ....ATT.CT T.T....... ...TC...A. ........TA .CA.GGA.CA .T|....T.GT.. .......... .....GTATC CA.CTT.... G.......GT
```

SA1619 **SA1621**

b

SA1619 **SA1621**

Figure 18. a) Variable nucleotide sites within closely linked ORPHANS SA1619 and SA1621. b) Representation of recombinational replacements in the region of SA1619 and SA1621.

3.3 RESULTS SUMMARY

There is no significant difference between functional constraint as measured by d_S/d_N and categorisation of genes according to function.

There is no significant difference in recombination rate as measured by the population-scaled recombination rate (ϱ) between categories or between pairwise category comparisons.

There is no significant difference in recombination rate as measured by the minimum number of recombination events (R_M) between categories except for housekeeping genes and OPRHANS.

There is no category of genes which is immune to recombination measure according to Bellerophon and Sawyer's Runs test.

There is no association between functional constraint, as measured by d_S/d_N,, and recombination rate as measured by the population-scaled recombination rate (ϱ), the minimum number of recombination events (R_M), Bellerophon and Sawyer's Runs Test.

The extent of recombination as detected by all methods is low enough to make the reconstruction of consensus phylogeny for *S. aureus* feasible.

A robust consensus phylogeny for *S. aureus* has been reconstructed with a high level of resolution between strains and good support by posterior probabilities.

A low level of recombination in this species is supported by the congruence of 34 out of 37 loci with the consensus phylogeny. Incongruence in the 3 loci is found to be a result of a paucity of informative sites and not due to extensive recombination.

There is no significant difference in d_S/d_N for all genes between the 3 population groups identified in the consensus tree.

There is no significant difference in recombination as measured by population-scaled recombination rate between the 3 population groups identified in the consensus tree.

There is no significant difference in SH score, representing phylogenetic reliability, between categories or between pairwise category comparisons.

The highest scoring loci are SA2439 (unknown function), *pbp*2 (cellular envelope and cellular processes) and SA1619 (ORPHAN).

Trees representing all genes of unknown function, cellular envelope and processes and the additional housekeeping gene have the highest SH scores against all other data.

There is no significant association between average pairwise diversity and SH score and a high level of variation in average pairwise diversity between the three highest scoring loci.

There is no significant association between phylogenetic reliability and recombination as measured by the population-scaled recombination rate (ϱ) or the minimum number of recombination events (R_M)

However, genes in which recombination is scored as present by Bellerophon score significantly higher by the SH test than those in which it is not detected.

There is no significant difference in SH score between presence and absence as detected by the Sawyer's runs test.

There is no association between SH score and distance from the origin of replication in this dataset.

Evidence from neighbouring loci suggests that the majority of recombination results in small localised replacements.

3.4 DISCUSSION

Intraspecific functional constraint

There is no evidence for differences in functional constraint represented by d_S/d_N between the samples of different categories of genes within this dataset. This observation is based solely on the sequences of 5-8 genes per functional category, representing an entire genome and with additional loci there could be a significant difference. King Jordan *et al*, exploited the existence of multiple complete genomes from single bacterial species (*C. pneumoniae, E. coli, H. pylori* and *N. meningitidis*) to examine the rates of evolution between different functional classes of genes. The d_S/d_N was determined for orthologous genes, grouped into 18 specific functional categories, and shared between strains of the same species. *C. pneumoniae* was the only species that did not show any evidence of significantly different rates of evolution for orthologs from different categories (Jordan *et al.*, 2002). The categorisation of the *S. aureus* genome into three defined categories (information pathways, housekeeping and cellular envelope and cellular processes) and 2 categorisations where function is not known (unknown function) and the absence of orthologs (ORPHANS) is a simplistic one. Such a categorisation, and sampling, is likely to under-represent the complexity of functional diversity within loci around the genome. However, in the absence of 30 intraspecific genomes, this sample per category represents a diverse selection of the genome typical for population-scaled studies such as phylogeny or for typing and so it is prudent to establish what such a categorisation encompasses. Despite the absence of a significant difference in d_S/d_N between categories we do observe differences in the extent of variation of d_S/d_N per category. ORPHANS have a narrow range of d_S/d_N between them (2.3 – 5.0) whereas the information pathways and cellular envelope and cellular processes genes have a much wider range of d_S/d_N within this sample (6.0- 125, 6.5 – 44.3). The housekeeping genes show an intermediate rate of sequence evolution (2.8 – 23.5). Interestingly, King Jordan *et al.*, also observe extensive variation within functional groups as well as between them (Jordan *et al.*, 2002). The measure of d_S/d_N however is a dynamic one. The absence of a significance difference in d_S/d_N between categories in the closely related *C. pneumoniae* genomes (Jordan *et al.*, 2002) most likely reflects the lack of time over which substitutions and purifying selection may occur. The effects of selection are not instantaneous and hence the d_S/d_N may change

over time due to a lag in the removal of slightly deleterious substitutions from the population (Rocha *et al.*, 2005, in press). Such an effect is rarely acknowledged and compromises the comparison of d_S/d_N between samples which represent different divergence times. In this study this variable is controlled for since each comparison is based upon the same strain set. We also observe differences in the frequency of synonymous substitutions although the difference between categories was not found to be significant ($p=0.364$). However, Jordan *et al.*, reported significant differences in pairwise genome comparisons between functional groups (Jordan *et al.*, 2002). Such an observation may reflect the dual action of some purifying selection on synonymous substitutions and some mutational biases. They also found to their surprise that one of the least conserved groups were those of DNA replication, recombination and repair. By our broader categorisation such loci would termed as information pathways and their result is consistent with the differences in the data observed here.

Good Housekeeping?

Housekeeping genes are typically recommended for intraspecific MLST schemes since they are assumed to represent the core genome and be under a purifying selection. However, no significant difference in d_S/d_N is found between housekeeping genes and genes of other functional categories. However the intermediate and relatively consistent rate of sequence evolution found within this dataset suggests that they provide an appropriate level of resolution for both short and long term epidemiological surveillance. However, atypical loci such as SA0008 are also found within the housekeeping genes. This loci is highly uninformative and would provide little resolution as part of a typing scheme. This gene encodes a housekeeping enzyme histidine ammonia lyase. There is no obvious functional reason why this particular loci should be more conserved than any other housekeeping gene. This gene is located close to the origin of replication. However, no significant association has been found between distance from the origin of replication and phylogenetic reliability.

MLST is often favoured as a typing scheme since the nucleotide data also lends itself to evolutionary analysis. An intermediate level of phylogenetic congruence was found for *S. aureus* MLST housekeeping loci and a lower rate of recombination to point mutation

frequency than found for other species (Feil *et al.*, 2003). Again no significant difference was found in recombination rates, as measured by several methods, between different categories of genes. The only significant difference was found between housekeeping genes and the ORPHANS regarding the Minimum number of recombination events (R_M). In this case the mean number of recombination events for the housekeeping genes and ORPHANS are 2.38 and 5.00 respectively. This indicates that there is slightly less recombination within housekeeping genes than ORPHANS. Estimates for the impact of recombination from MLST data for *S. aureus* can be considered representative of the genome. However, inspection of polymorphic sites and the presence of long branches for some strains in the tree for the housekeeping gene *hutI* revealed the incorporation of highly divergent (14%) and yet largely synonymous recombinational replacement. Although a donor for the sequence cannot be identified the level of divergence is highly indicative of interspecific homologous recombination. The maintenance of allelic integrity (by being synonymous) has facilitated the rise of this mosaic to observable frequency in the population. Such an event in the evolutionary history of a single loci compromises accurate characterisation and phylogeny and highlights the importance of the use of multiple loci for both typing and phylogenetic purposes.

Information Pathway genes

Essential genes such as those involved in protein synthesis, DNA replication and repair are represented within the information pathway gene category and are more likely to be conserved through speciation to be represented in a groups of orthologs between taxa. However, does the conservation of such loci confer immunity to recombination? Firstly, we find no evidence for an increased functional constraint within the information genes sampled. However, nonsynonymous substitutions are absent in two of the seven loci and no d_S/d_N can be calculated for them. The paucity of informative sites in these two loci, *sarA* and *tufA* provides little resolution resulting in no more similarity to the consensus tree than trees of random topology. *SarA* has only one polymorphic site present in only one strain. Hughes and Friedman report an analysis of the patterns of substitution in orthologous genes of 5 *S. aureus* genomes. They find 108 orthologs within sister pairs of genomes (MW2, mss476: ST1 and N315 and Mu50) and a more divergent genotype (mrsa252: ST36) which were identical synonymous sites (Hughes & Friedman, 2005). The

only known form of selection affecting synonymous sites is selection on synonymous codon usage but this selection is likely to be purifying and thus reduce the rate of synonymous substitution rather than enhance it (Sharp, 1991). An alternative hypothesis to explain this observation may be the frequent recombination within these loci whereby a positive feedback loop purges variation that has arisen by point mutation (Cohan, 1995). Detecting recombination in cases where there is almost complete identity would be impossible. Two of the highest d_S/d_N values found within the dataset fall within this category representing strong purifying selection. Although there is no significant difference between functional constraints in this sample these results suggest that a stronger functional constraint may be found with additional data. However, we also find no difference between functional constraint and the extent of recombination within loci. Despite the strength of purifying selection observed some of these loci, recombination is still detected within them. SA0189 is classified as a probable type I restriction enzyme restriction chain and it has the highest calculated population-scaled recombination rate (ϱ) of all genes in the dataset. Evidence for recombination within *dnaC* is provided by all the recombination tests used in this study (Table 5). This gene encodes an essential DNA helicase. The fact that this loci has a high d_S/d_N of 125 representing synonymous variation is remarkable. Despite the essentiality of its function as a replicative DNA helicase required for DNA replication (Bruck & O'Donnell, 2000) this gene has not remained conserved like *sarA* and *tufA*, homologous recombination has occurred resulting in synonymous changes. Thus, recombination is clearly compatible with selective constraint in the case where replacements are synonymous. Furthermore, the gene *agrC* forms part of the *agr* (accessory gene regulator) locus which controls the production of exoproteins implicated in virulence (Morfeldt *et al.*, 1988; Peng *et al.*, 1988; Recsei *et al.*, 1986). Indeed the pattern of action for this locus is complex, upregulating certain extracellular toxins and enzymes expressed postexponentially and repressing some exponential phase surface components (Novick & Muir, 1999). *AgrB* triggers the *agrD* encoding of a precursor which interacts with *agrC* resulting in the activation of *agrB* in a two component sensory transduction system. *AgrB* in turn upregulates transcription amplifying the response and initiating the production of a novel effector (Janzon & Arvidson, 1990; Novick *et al.*, 1993). The polymorphic nature of this locus and its existence in four distinct genetic types in *S. aureus* (Jarraud *et al.*, 2000; Ji *et al.*, 1997) and other staphylococcal species (Dufour *et*

al., 2002) has been shown. The hypervariability extents from the C-terminal of *agrB*, across *agrD* and into the N terminal end of *agrC*. A maximum likelihood tree generated from internal sequence of *agrC* is shown in Figure 19. The groupings on the tree clearly shows the 4 different agr types for *S. aureus*. The colour of the bracket indicates the population Group the strains enclosed represent. Although some STs do not always have the same *agr* type (Peacock *et al.*, 2002) this cannot be represented within the scope of the data collected here as the sample of identical STs is too small. However, we can clearly see that for the *agrC* data collected here that H512 ST1 from Group2 is grouped with strains from Group1 of the *S. aureus* population. Since the population groupings are based on the sequences of 38 loci we can confidently state that this is evidence for recombination within the *agr* locus.

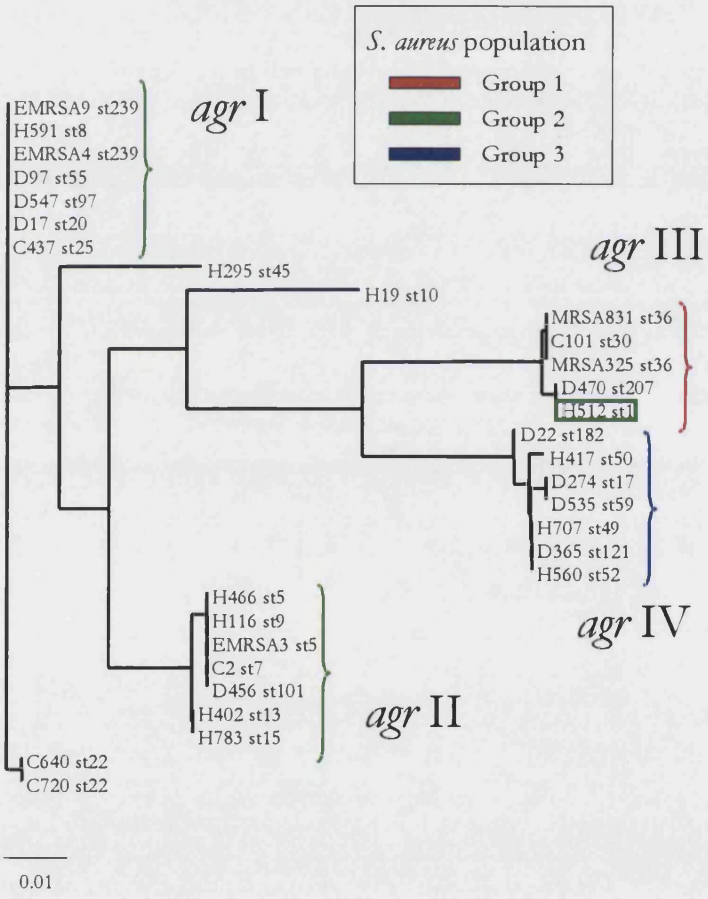


Figure 19. *AgrC* maximum likelihood tree with agr groups (*agr* typing data adapted from (Peacock *et al.*, 2002).

Agr Group is further shown to be an unstable characteristic of strain lineage and subject to recombination by a study by Robinson *et al.*, (submitted). In their study they find incongruence between *agr* trees and a combined dataset representing 7 MLST housekeeping genes and 7 staphylococcal adhesin surface (sas) protein encoding genes. In the case of *agrC*, not all the variation is synonymous, so despite complex interactions non-synonymous variation has still accumulated, and by recombination. The evidence for recombination within *dnaC* and *agrC* is of particular interest in light of the 'complexity hypothesis' which suggests that recombination in information pathway genes is less likely due to the complexity of the interactions between such genes. *AgrC* is part of a complex interacting regulation system and this has not conferred immunity to recombination. Such a barrier of complexity may be removed with the transfer of the entire operon. This potentially creates a misleading situation since recombination events which encompass the entire sequence subject to analysis are more easily over looked. We have observed both synonymous and nonsynonymous replacements within information pathway genes providing little evidence that the 'complexity' hypothesis applies at the intraspecific level. Many of the information pathway genes score poorly by the SH test with *luxS* scoring highest falling in 10th place behind genes from all other categories (Table 8). The consensus tree generated from the concatenated information pathway data does not score as well as consensus trees for genes of unknown function, cellular envelope and cellular processes genes or housekeeping genes. Although, information pathway genes fare better than ORPHANS and considerably better than the MLST housekeeping genes alone. From every perspective in which we have analysed the data we find no evidence that information pathway genes are immune to the effects of recombination or more phylogenetically reliable using *S. aureus* as a model to test this hypothesis.

Staphylococcus aureus phylogeny

Recombination within diverse gene types has occurred in the evolutionary history of *S. aureus* and several examples have been highlighted both through the inspection of polymorphism distribution, the topologies of individual loci phylogenies and the implementation of several methods for the detection of recombination. However, this recombination is not extensive within the sample of 38 genes analysed here and phylogenetic signal has been maintained. A robust and well supported phylogeny is

presented for lineages representing the natural population of *S. aureus* in this study. Melles *et al.*, report the *S. aureus* population structure from amplified fragment length polymorphism (AFLP) data (Melles *et al.*, 2004). This method scans for polymorphism in restriction sites and flanking nucleotides documenting the contribution of both accessory and core loci polymorphisms. Their study characterises variation in a large number of non-clinical isolates (n=829) and a range of *S. aureus* disease isolates, including some MRSA (n=235). Clustering analysis reveals the presence of three major groups and two subgroups (Figure 20).

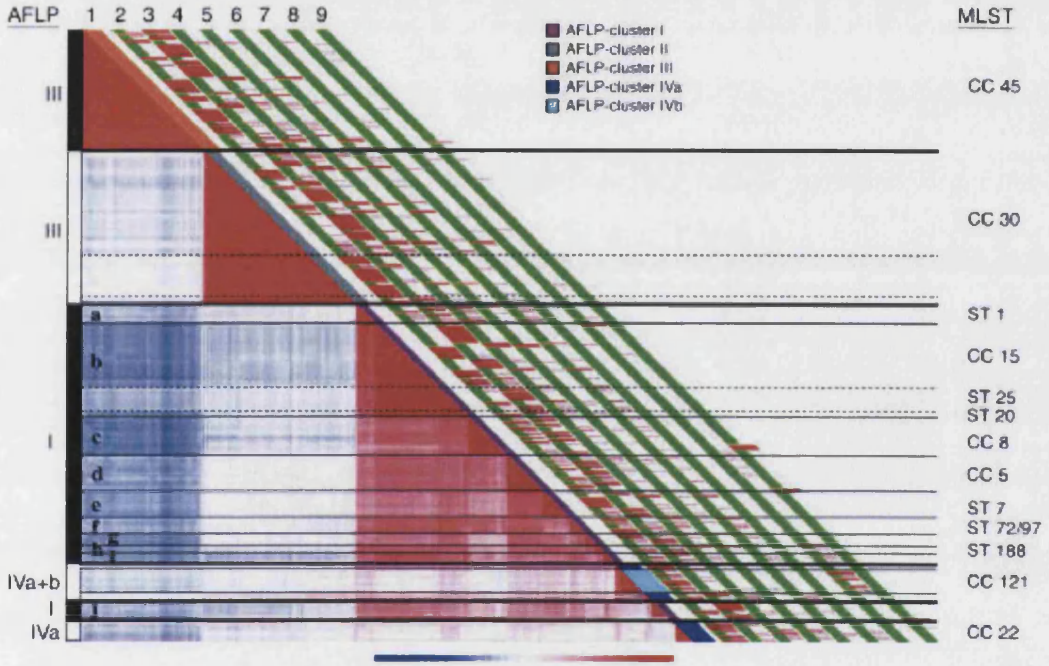


Figure 20. Cluster analysis of the 1,056 *S. aureus* taken from Melles et al., 2004.

Of relevance in this figure are the MLST data shown on the right side of the figure and the 5 AFLP groups indicated by the black and white bar on the left. The cells in the correlation visualization are colored by Pearson's correlation coefficient values with deeper colors indicating higher positive (red) or negative (blue) correlations.

Group I in the Melles study (Figure 20) correlates to the Group2 definition in this thesis which includes representatives of clonal complexes 8, 15, 5 and 25. The two other major Groups, II and III, are represented by CC30 and CC45 respectively. These two complexes represent divergent lineages in the *S. aureus* population. However, the isolates that are included in the Melles study are unlikely to include ST207 and ST10 which also represent diverse but rare lineages of *S. aureus* but for which each there is only a single isolate. The presence of these lineages in the phylogeny presented in this thesis results in the inconsistency of the groupings between these two studies. If these further diverse lineages had been included in the Melles study they may have also considered grouping these lineages together rather than further divisions within the population – i.e. 7 groups. The Melles study also identifies CC22 as an individual subgroup. CC22 represents another divergent lineage in Group1 of the population as identified in this study. Unlike divergent lineage CC45 however, this particular lineage has not been assigned as a major Group but as subgroup (IVa) due to the fewer isolates representing this clonal complex. Subgroup IVb of the Melles study represents Group3 of the population in this study. The clustering of strains within both studies is consistent. The differences exist in the identification of population divisions. This also illustrates the impact of sampling effects even where the sample size is huge and most diversity is represented but how this can affect our interpretation of the population structure.

Phylogenetic reliability in this study is used as a measure of how well the evolutionary history of the individual loci reflects the evolutionary history of the organism (as represented by all other data). Table 8 shows the phylogenetic reliability of each individual gene and in the lower part of this table we observe several MLST genes which have previously been reported as the least congruent of the MLST genes (Feil *et al.*, 2003). This result is therefore consistent with those findings. The only noncongruent loci compared to the data for all other genes are SA0008, *sarA* and *tufA* (Table 7). In Figure 12 these 3 genes represent the lowest average pairwise diversity and therefore this incongruence is a result of a paucity of informative sites resulting in poor resolution. Yet the remaining 34 genes included in the phylogeny were found to be significantly more similar to the consensus phylogeny than genes of random topology. This level of phylogenetic reliability and the agreement of lineage assortment by AFLP in the Melles study supports the use of

concatenation of multiple loci to overcome the effects of recombination in individual loci. Indeed, Robinson and Enright report the existence of two large replacements of ~244 and 557 kb within *S. aureus* lineages. The recombinant lineage ST239 and parental lineages ST8 (Group2) and ST30 (Group1) are included in this study. ST239 is an SLV of ST8. The recombination replacement spans SA2339, includes the origin of replication and ends within SA0318. Within this region ST239 has almost complete identity to ST30. However, in remainder of the genome this lineage resembles its ancestor by descent ST8 (Robinson & Enright, 2004). The relationships between these lineages within individual gene trees change according to the genomic location of the gene, consistent with the existence of this replacement. In the consensus phylogeny generated for *S. aureus* ST239 remains closely associated to its ancestor ST8 despite the inclusion of 17 loci which fall within the region of the replacement. This again supports the use of concatenation as a phylogenetic method to overcome the impact of recombination not only at single loci but within genomic regions provided that a diverse genomic location is sampled.

In identifying individual loci that best represented the consensus phylogeny we were unable to identify any parameters that could systematically predict the phylogenetic reliability of single genes. A gene of unknown function, SA2439, was found to best represent the consensus phylogeny. This gene was not annotated in the N315 genome. However, a blastn search reveals that this gene encodes (*staphylococcus aureus* surface protein) sasF. This gene has an LPXTG (although it varies at one residue: LPKAG) motif and is anchored by the action of sortase in the same way as the characterised surface proteins of *S. aureus* (Roche *et al.*, 2003). This seems surprising that a gene encoding a surface exposed protein would best reflect the phylogeny of the species. However, this supports the use of *sas* genes in phylogenetic analysis in the study of *S. aureus* genomic replacements and the evolutionary history of and *agr* types (Robinson & Enright, 2004). This gene has not been immune to recombination (Table 5), nor is it the most diverse gene (which may generate long branches comparable to those in a consensus phylogeny of 37 genes). However, this gene alone is not recommended as an indicator of *S. aureus* phylogeny. The merits of multilocus phylogeny have previously been discussed. As concatenated categories the genes of unknown function are the closest fit to the consensus phylogeny. Interestingly, the fact that these genes are uncharacterised suggests that there is

little phylogenetic pattern within categories and that phylogenetic reliability is a much more stochastic process. Again, unpredictably, the cellular envelope and cellular process also score highly. The additional housekeeping genes score much higher than the MLST housekeeping genes. Three of the additional housekeeping genes are represented by sequences which are twice the size of a typical MLST gene. These, not surprisingly, are the three highest scoring housekeeping loci. Several MLST housekeeping genes score poorly by the SH test as a result of poor resolution due to a paucity of informative sites. This indicates that although ~450 bp fragments are suitable for typing a larger fragment may enable more meaningful phylogenetic analysis of the sequence data. These results, along with the evidence for recombination and poor levels of phylogenetic reliability within information pathway genes, challenge our perceptions over which gene types are most appropriate for bacterial phylogeny and require that the criteria for the interpretation of phylogenetic markers, and MLST genes is re-evaluated.

CHAPTER FOUR

THE DISTRIBUTION OF SDR GENES IN THE NATURAL POPULATION OF *STAPHYLOCOCCUS AUREUS*

CHAPTER FOUR: DISTRIBUTION OF SDR GENES

4.1 INTRODUCTION

The sdrE gene encodes a putative surface adhesin involved in the activation of human platelet aggregation. This gene has been observed in two allelic forms, the sdrE form of which has been associated with invasive disease isolates (Peacock *et al.*, 2002).

The aim of this Chapter is to analyse the distribution of the *sdrE* locus between disease and carriage isolates and within lineages and clonal complexes of the natural population of *S. aureus*.

The phylogenetic tree in Figure 1 illustrates the grouping of the 3 *S. aureus sdr* genes. *Bbp* from strain mrsa252 clearly groups with the *sdrE* genes from strains mssa476, COL and 8325. We can also observe the absence of *sdrD* in strain mrsa252 (Holden *et al.*, 2004; Kuroda *et al.*, 2001).

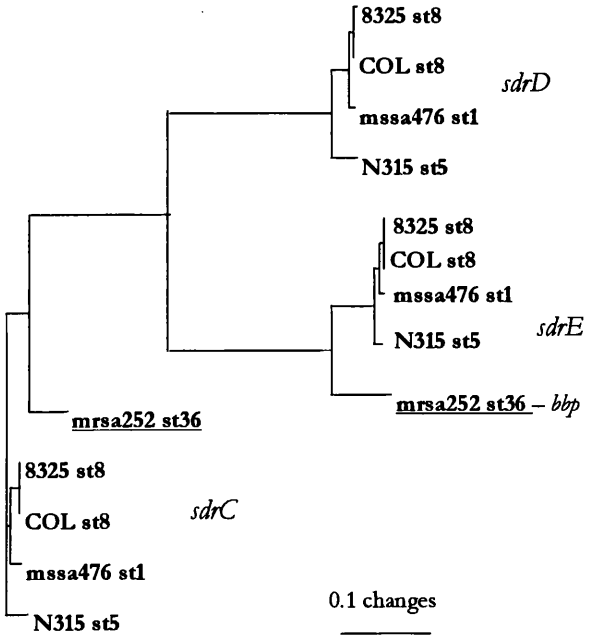


Figure 1. Neighbour-joining tree of staphylococcal *sdr* genes from complete genome sequenced strains. Strains with only 2 genes at this locus are underlined.

s*drE* and *bbp* specific primers are used on both the Oxford and Nottingham collections representing disease and carriage *S. aureus* isolates. Bacterial strains are as described in Chapter 2, page 41. The *sdrE* and *bbp* specific primer strategy is shown below in Figure 2 (Peacock *et al.*, 2002). The resulting amplicons for *sdrE* and *bbp* alleles are 766 bp and 1054 bp respectively and are visualised by gel electrophoresis with ethidium bromide staining (Chapter 2, page 44).

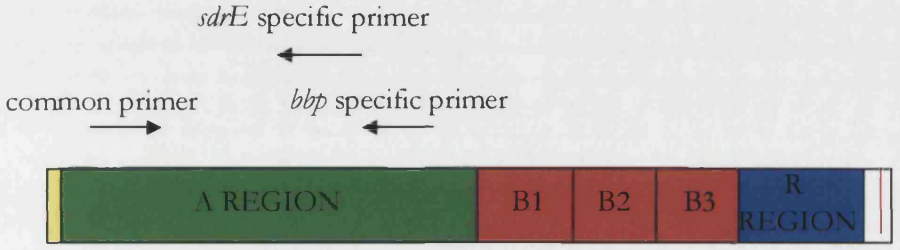


Figure 2. *SdrE* and *bbp* PCR strategy.

Banding patterns are recorded and negative results (indicating the absence of the *sdrE* locus) were further validated. The integrity of the genomic DNA was confirmed by the amplification of an MLST gene. These data are then used to assess the distribution of the *sdrE* locus and both its allelic types (*sdrE* and *bbp*) within the context of both clonal complexes and the relationships between lineages resolved in Chapter 3, page 90. *SdrD* is directly downstream of *sdrE* and specific primers are also used to determine its distribution in the population relative to the *sdrE* locus.

All primer sequences can be found in the Appendix B1.

4.2 RESULTS

4.2.1 Frequencies of *sdrE* and *bbp*

The presence of either the *sdrE* or *bbp* amplicon represents the presence of the *sdrE* locus. The frequencies for the *sdrE* locus and both allelic types are given in Figure 3. The *sdrE* locus was present in 86% of all 485 strains included in this study. The *sdrE* type allele is found within 55% of all strains whilst the *bbp* allele accounted for only 30% of these strains. Raw data for the Oxford collection can be found in appendix B2. Raw data for the Nottingham collection can be found in appendix B3.

a)

Collection	isolation source	n=	*sdrE* locus
Oxford	hospital acquired disease	91	84
Oxford	community acquired disease	60	56
	total disease isolates	**151**	**140**
Oxford	asymptomatic carriage	174	154
Nottingham	asymptomatic carriage	160	123
	total carriage isolates	**334**	**277**

b)

Collection	isolation source	n=	*sdrE*
Oxford	hospital acquired disease	91	50
Oxford	community acquired disease	60	40
	total disease isolates	**151**	**90**
Oxford	asymptomatic carriage	174	92
Nottingham	asymptomatic carriage	160	85
	total carriage isolates	**334**	**177**

c)

Collection	isolation source	n=	*bbp*
Oxford	hospital acquired disease	91	33
Oxford	community acquired disease	60	16
	total disease isolates	**151**	**49**
Oxford	asymptomatic carriage	174	60
Nottingham	asymptomatic carriage	160	38
	total carriage isolates	**334**	**98**

Figure 3. a) The frequency of the *sdrE* locus within disease and carriage isolates. b) the frequency of the *sdrE* type allele within disease and carriage isolates c) the frequency of the *bbp* allele within disease and carriage isolates.

The discrepancy between the number of *sdrE/bbp* alleles and number of isolates in which the *sdrE* locus is found to be present is due to the presence of both alleles within some isolates.

The frequencies of *sdrE* within disease and carriage isolates

The chi-squared test of association was used to test the hypotheses that there is no significant difference in the frequency of the *sdrE* locus or its allelic variants within isolates from asymptomatic carriage and from invasive disease.

sdrE locus

	PRESENT	ABSENT
DISEASE	140	11
CARRIAGE	277	57

Chi-squared = 8.253
df = 1
p= 0.004

sdrE allele

	PRESENT	ABSENT
DISEASE	90	61
CARRIAGE	177	157

Chi-squared = 1.835
df = 1
p= 0.175

bbp allele

	PRESENT	ABSENT
DISEASE	49	102
CARRIAGE	98	236

Chi-squared = 0.476
df = 1
p= 0.490

Table 1. Chi-squared test of association for *sdrE* and disease and carriage isolates.

SdrE is significantly more frequent in isolates from disease compared to carriage p> 0.01. However, when the *sdrE* locus is differentiated into *sdrE* and *bbp* alleles no significance is found for different allele types.

4.2.2 Distribution of *sdrE* locus in the population

The *sdrE* locus is found within all three groups of the population and within most lineages (Figure 4). The most parsimonious explanation for the absence of this locus is three independent losses in ST13, ST10 and ST182. These three lineages represent each of the three major population groups. Figure 5 shows the distribution of the *sdrE* locus within the clonal complexes of *S. aureus*. The strains represented within these clonal complexes are those from the Oxfordshire collection (Chapter 2, page 41). Although the *sdrE* locus is present within all the clonal complexes here there have been losses of this locus within at least one isolate representing the ancestral ST of 6 of the clonal complexes, and within SLV and DLVs (Table 2).

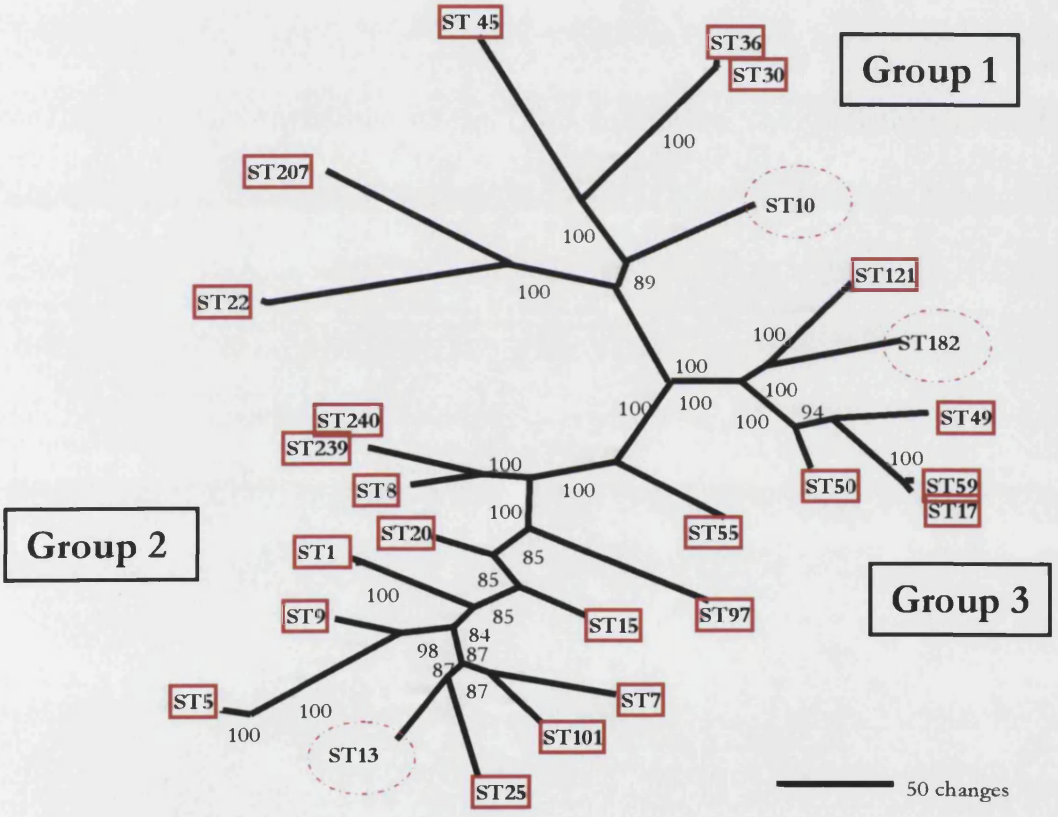


Figure 4. Distribution of the *sdrE* locus within *S. aureus* lineages. The presence of the *sdrE* locus is denoted by the red boxes. Absence is denoted by the red dashed circles.

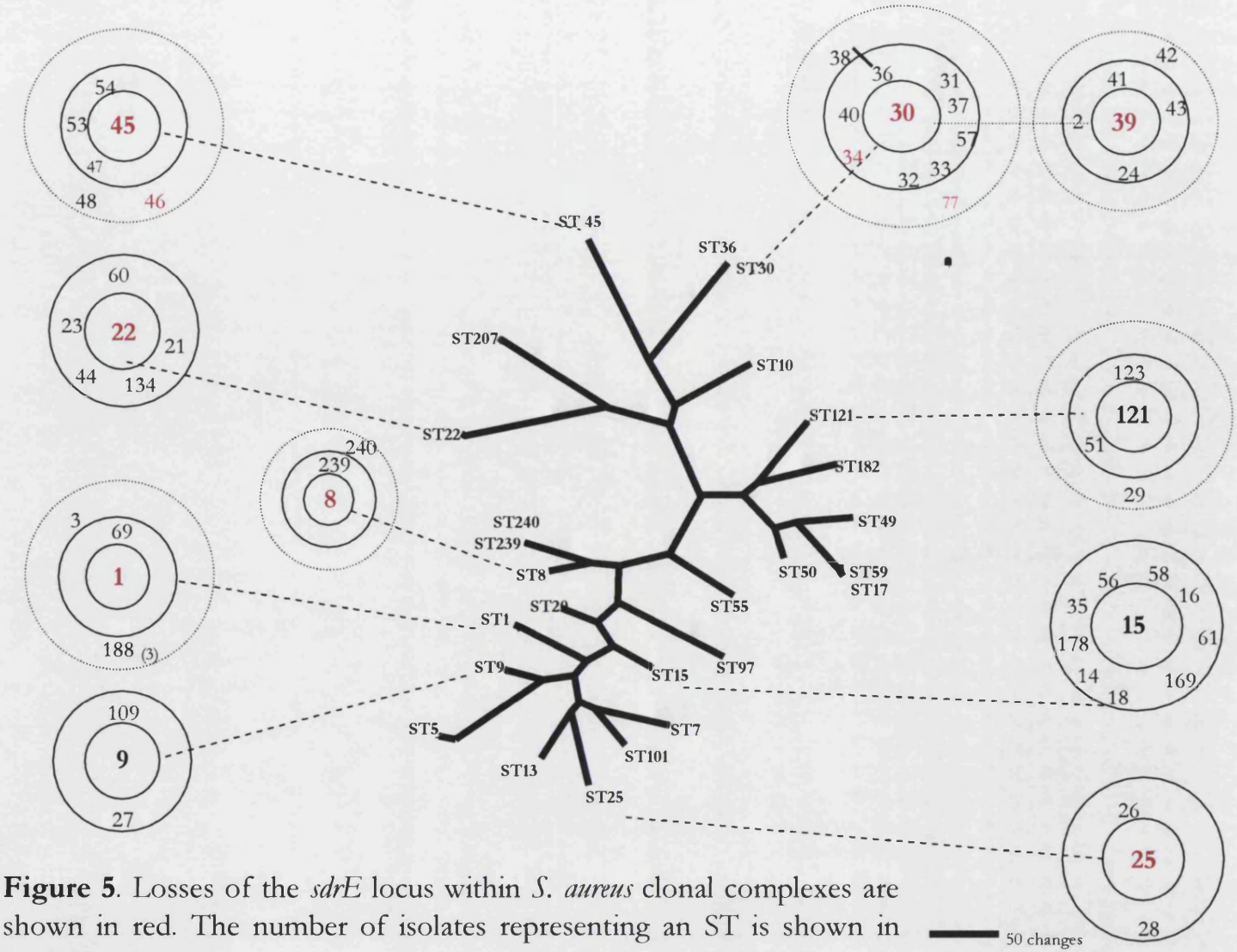


Figure 5. Losses of the *sdrE* locus within *S. aureus* clonal complexes are shown in red. The number of isolates representing an ST is shown in parentheses

GROUP	CC	ST	LOSSES	(OF TOTAL ISOLATES)
1	45	45	2	13
1	22	22	4	20
1	30/39	39	3	19
1	30/39	30	3	48
1	30/39	77	1	1
1	30/39	34	2	12
2	8	8	1	16
2	1	1	2	11
2	25	25	1	21

Table 2. Losses of the *sdrE* locus within *S. aureus* clonal complexes.

These losses are shown in the context of the clonal complexes and the *S. aureus* phylogeny in Figure 5.

The '*bbp*' type allele is observed within all three groups of the population, within three lineages: the closely related ST30 and ST36, ST121 and also the closely related lineages of ST7 and ST101 (Figure 6). The resolution of branches clearly demonstrates that this allele has arisen within all three groups by recombinational transfer and not by descent. This presumably occurred in the ancestral sequence of ST30 and ST36, in the ancestral sequence of ST7 and of ST101. The *bbp* allele may have arisen prior to the diversification of ST121 and ST182, then to be lost in ST182. Alternatively, it may have arisen in ST121 alone, post diversification from ST182. Figure 7 shows the distribution of the *bbp* allele within the clonal complexes of *S. aureus*. This allele is found to be predominant within the clonal complexes CC30/39 and cc121. It is also observed atypically within an SLV and a DLV of the ST45 clonal complex and an SLV of the ST9 clonal complex, the *sdrE* allele is found within all other isolates of these complexes.

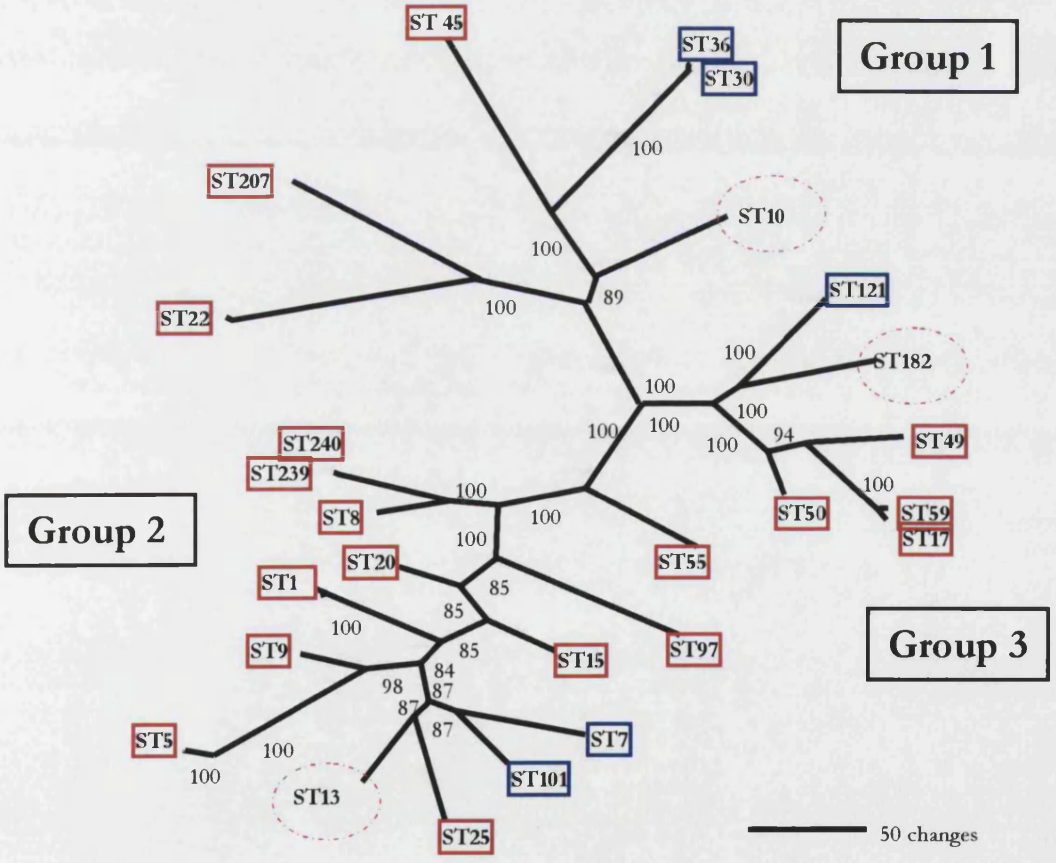


Figure 6. Distribution of allelic types at the *sdrE* locus. The *bbp* allelic form of the *sdrE* locus is denoted by the blue boxes.

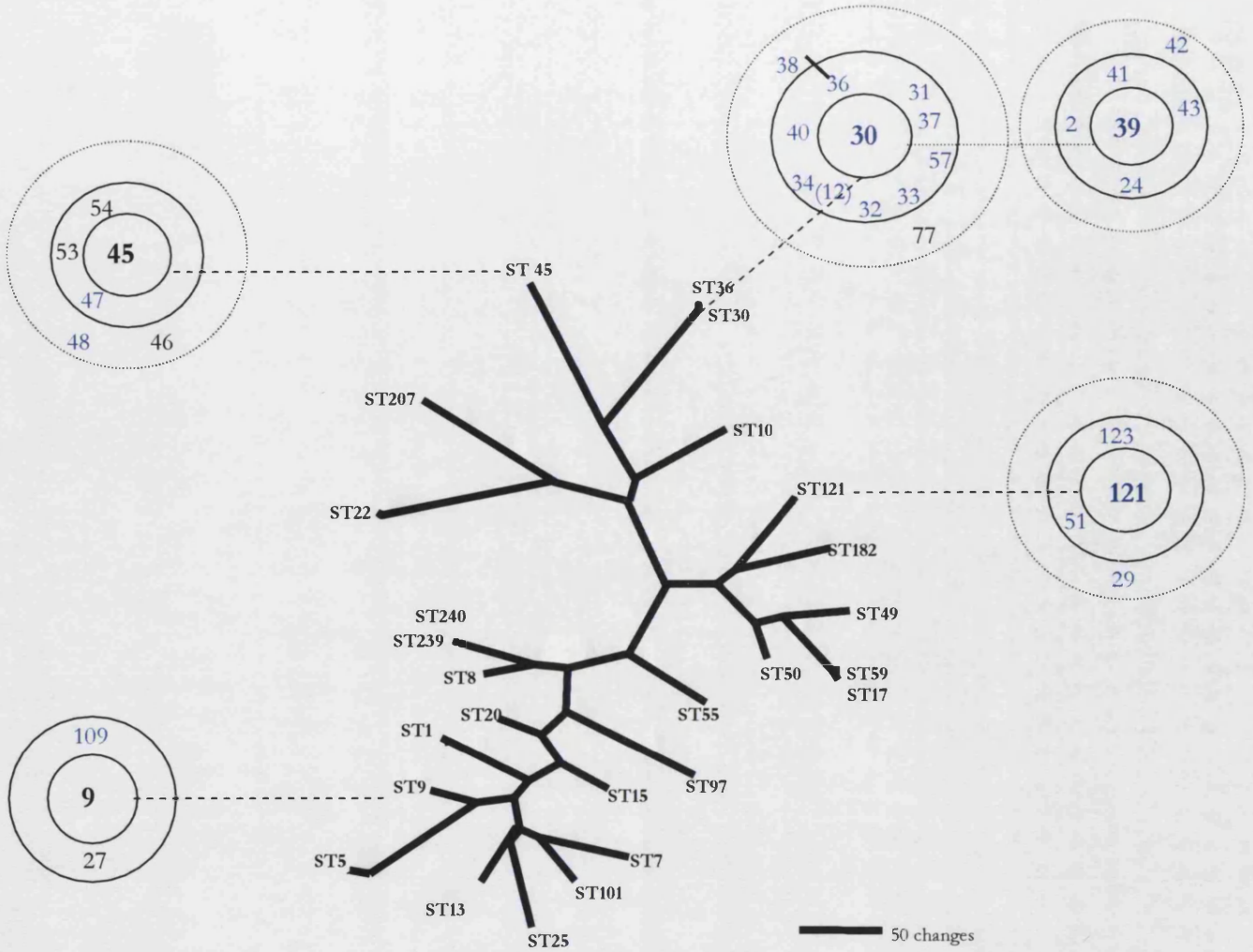


Figure 7. Distribution of *bbp* allele within *S. aureus* clonal complexes. Presence of *bbp* is shown in blue

Some strains gave a positive result for both *sdrE* and the '*bbp*' type allele. These were found within the large cc30/39 clonal complex within one isolate of the ancestral founder ST30 and 2 isolates of ancestral founder ST39 and also within 4 isolates of ST36, the EMRSA-16 strain (Figure 8). This result was validated by the use of a further PCR to determine the arrangement of the 2 *sdrE* genes relative to each other. Primer sequences and the potential orientations of these two can be found within Appendix B. The *bbp* allele was found to lie directly downstream of the *sdrE* allele (Figure 8).

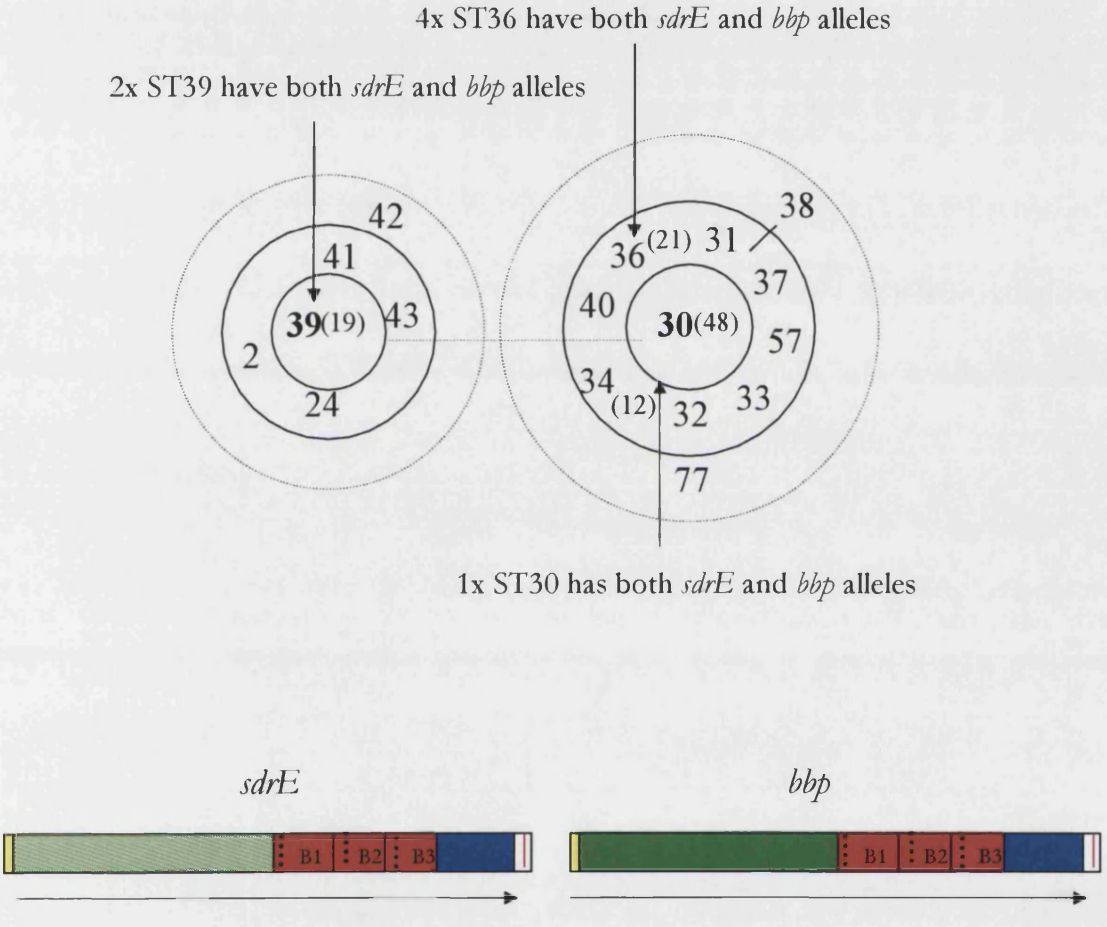


Figure 8. Presence and arrangement of 2 *sdrE* genes within CC30/39.

4.2.3 Distribution of *sdrD* in the population

The presence of *sdrD* was assayed by PCR to determine its distribution in the lineages identified within the population framework and is shown in Figure 9. The *sdrD* locus is present in all lineages in Group 2 of the population with the exception of ST13 and ST97. It is present in only 2 lineages outside of Group 2, ST17 and ST22 (Figure 9). The most parsimonious explanation for this distribution is the loss of the *sdrD* locus in the common ancestor of Group 1 and Group 3, after diversification from Group 2. *SdrD* has then been lost in ST13 and ST97 of Group 2 and then gained in lineages ST22 (Group 1) and ST17 (Group 3).

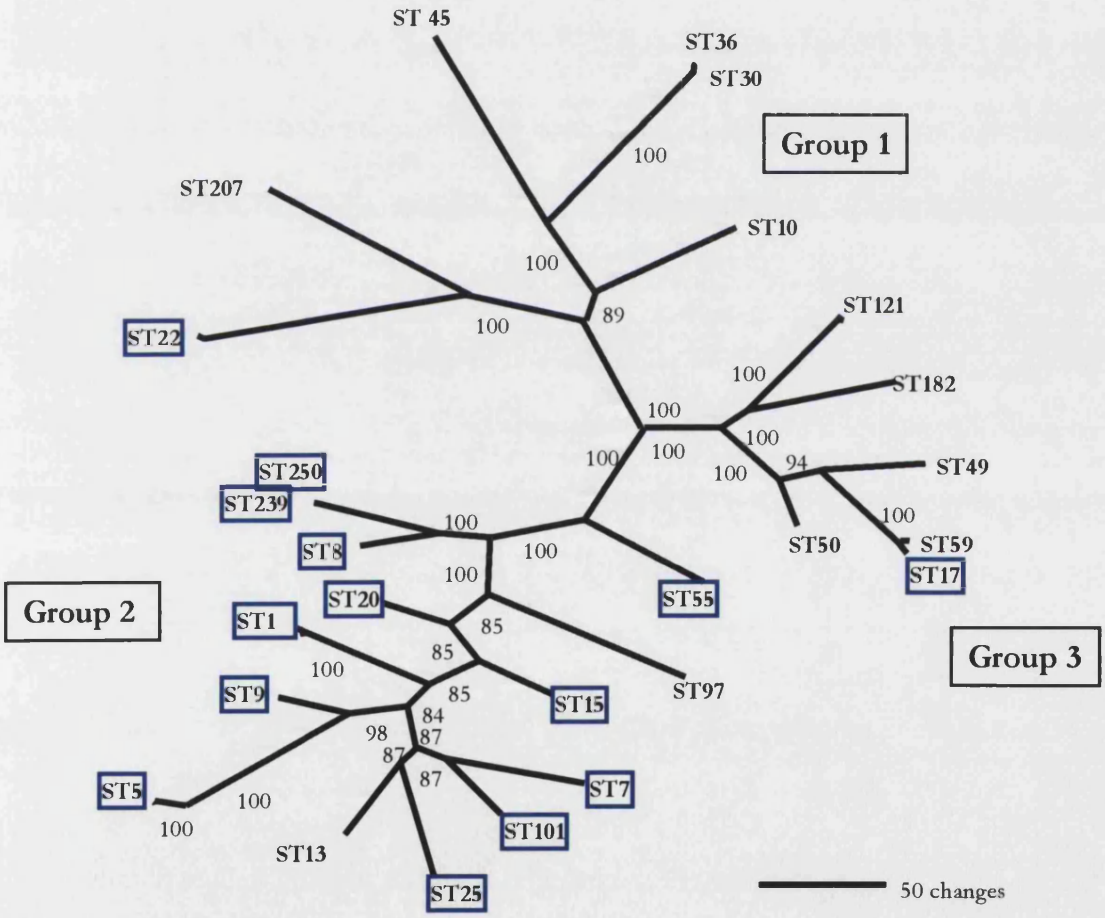


Figure 9. Distribution of *sdrD* locus within *S. aureus* lineages.

4.3 DISCUSSION

The study of Peacock *et al.* reports a significant association of the *sdrE* type allele with invasive disease (Peacock *et al.*, 2002). Although we cannot confirm such as association there is a larger proportion of *sdrE* allelic types associated with the disease isolates than with carriage strains. The inclusion of many more carriage isolates in this study may have diluted such an effect since it is possible that an isolate from a healthy carrier could also have virulence potential with a change in human-associated and stochastic factors. However, we do find a significant association of the presence of the *sdrE* locus and disease isolates. The majority of cases of severe staphylococcal disease are unlikely to be the result of a single virulence factor but a multifactorial effect. Peacock *et al.* demonstrated this with a positive increase in the percentage of isolates representing disease with the accumulation of factors associated with staphylococcal virulence (Peacock *et al.*, 2002).

The well resolved phylogeny generated in Chapter 3 allows the analysis of the distribution of the *sdrE* locus within the context of the population framework. With such a framework we can reveal that the *sdrE* locus is conserved in most lineages and the branching orders confirm that this locus has been lost in three lineages, representing the three population groups, in three independent events. Within more recent evolutionary history we also observe the loss of the *sdrE* locus within both clonal complexes and within strains of the same ST. In other words the loss of the *sdrE* locus has occurred both over the longer and shorter evolutionary history of this species. The *sdrE* allelic form of this locus is the most common form representing 62% of strains where the locus is present.

The *bbp* allelic form has arisen in lineages within each of the 3 groups of the population to become the dominant allelic form in CC30/39 and 121 and also in lineages ST101 and ST7. These represent much older events in the history of the population. More recently the *bbp* allelic form has also arisen within some isolates of 2 further clonal complexes in which *sdrE is* typically predominant. The presence of the locus within the majority of the population and the distribution of the *bbp* allele clearly demonstrates that recombination involving the *bbp* allelic form has also occurred also over both the longer and shorter evolutionary history of this species. The presence of both allelic types in some isolates of

the CC30/39 is indicative of the lateral transfer and incorporation of a second *sdrE* locus, of the *sdrE* allelic form, since this clonal complex is predominantly represented by the *bbp* allele. The distribution of the *sdrD* locus does not mirror that of the *sdrE* locus. It is found that *sdrD* is absent in ST97 where the *sdrE* locus is present. However, both the *sdrE* and *sdrD* loci are found to be absent in lineages ST13, ST10 and ST182. In contrast , *sdrC* is known to be present within all *S. aureus* lineages (Peacock *et al.*, 2002). The tandem arrangement and presence of serine-apartate repeats of the 3 sdr genes facilitates the deletion of a single locus, *sdrD* or *sdrE*, or both loci simultaneously. The observed distribution of the *sdrD* locus in Figure 9 suggests that *sdrD* was lost in the ancestral sequence that subsequently diversified to create population Groups 1 and 3 to be regained in the lineage of ST22 and more recently in ST17 (and not the very closely related ST59). We here find evidence for the continual loss and gain of both the *sdrD* and *sdrE* loci. We also present evidence for allelic replacement, presumably by recombination of both *sdrE* allelic forms over the evolutionary history of *S. aureus*.

In light of this the intraspecific transfer and allelic replacement of virulence-associated loci, the substitution of allelic variants may have implications in the development of virulence potential of *S. aureus* isolates, effective disease management and the development of vaccines. It should be noted that PCR based assays are limited by the specificity of primer design. However the utilisation of micro array technology to determine gene content will facilitate further analysis of differences in the distribution of virulence-associated loci in *S. aureus*.

The extent of the variation found within and between the *sdrE* alleles and the functional implications is presented, and will be discussed further, in Chapter 5.

CHAPTER FIVE

THE CHARACTERISATION AND FUNCTIONAL IMPLICATIONS OF VARIATION AT THE *SDRE* LOCUS

CHAPTER FIVE: CHARACTERISATION AND FUNCTIONAL IMPLICATIONS OF VARIATION AT THE SDRE LOCUS

5.1 INTRODUCTION

In Chapter 4 the presence of the *sdrE* locus was shown to be significantly associated with isolates from disease. The *sdrE* gene is a putative surface adhesin with a role in the activation of platelet aggregation. This gene exists in two divergent allelic forms. Evidence for gene transfer and allelic replacement by homologous recombination is also presented in Chapter 4.

There are several aims in this Chapter:

1. The characterisation and localisation of molecular variation within the *sdrE* and *bbp* allelic types.

2. To further investigate the impact of homologous recombination at this locus.

3. To test the conservation of function between proteins from diverse *sdrE* alleles.

Strain	ST	Epidemiology	Resistance Profile	*sdrE* locus profile	Gene sequencing	A region sequencing	Aggregation experiments
H512	1	HA-disease	MSSA	*sdrE*		YES	
C521	5	CA-disease	MSSA	*sdrE*	YES	YES	
D10	5	Carriage	MSSA	*sdrE*	YES	YES	
C2	7	CA-disease	MSSA	*bbp*		YES	
H591	8	HA-disease	MSSA	*sdrE*		YES	
H116	9	HA-disease	MSSA	*sdrE*		YES	
H783	15	HA-disease	MSSA	*sdrE*		YES	
D274	17	Carriage	MSSA	*sdrE*		YES	
D17	20	Carriage	MSSA	*sdrE*		YES	
B426	22	Carriage	MSSA	*sdrE*	YES	YES	
C13	22	CA-disease	MSSA	*sdrE*	YES	YES	
B87	25	Carriage	MSSA	*sdrE*	YES	YES	
C101	30	CA-disease	MSSA	*bbp*	YES	YES	YES
D363	30	Carriage	MSSA	*bbp*	YES	YES	
B504	30	Carriage	MSSA	*bbp*	YES	YES	
H513	30	HA-disease	MSSA	*bbp*	YES	YES	
B226	30	Carriage	MSSA	*bbp*	YES	YES	
C767	39	CA-disease	MSSA	*sdrE* and *bbp*		*sdrE* YES	YES
C730	45	CA-disease	MSSA	*sdrE*	YES	YES	YES
H707	49	HA-disease	MSSA	*sdrE*		YES	
H417	50	HA-disease	MSSA	*sdrE*		YES	
D97	55	Carriage	MSSA	*sdrE*		YES	
D535	59	Carriage	MSSA	*sdrE*		YES	
D547	97	Carriage	MSSA	*sdrE*		YES	
D346	101	Carriage	MSSA	*bbp*		YES	
D365	121	Carriage	MSSA	*bbp*		YES	
D470	207	Carriage	MSSA	*sdrE*		YES	
EMRSA4*	239	EMRSA type strain	MRSA	*sdrE*		YES	
EMRSA9*	240	EMRSA type strain	MRSA	*sdrE*		YES	

Table 1. Bacterial Strains used in Chapter 5 sequencing and experiments.

5.2 RESULTS

5.2.1 Localised variation between *sdrE* and *bbp* alleles

Eleven *sdrE* genes (including *bbp* alleles) of approximately 3.5 kb were sequenced in order to examine diversity within and between sequence types and to compare alleles between isolates from disease and those from asymptomatic carriage. Strains of the same sequence type tended to have the same size *sdrE* genes although there were variations in the size of *bbp* alleles within ST30. *SdrE* genes ranged from 3324 bp to 3486 bp (Table 2). The differences in sizes are found to be attributable to varying numbers of serine-aspartate (SD) repeats found within the R region of the sdrE protein. Previous studies on the R region in the clumping factor proteins confirm that this region acts as a stalk through the cell wall allowing full exposure of the biologically active A region (McDevitt & Foster, 1995). The sizes of all other domains of the *sdrE* gene were found to be conserved. The reconstruction of the *sdrE* nucleotide sequences and other family members *sdrC* and *sdrD* in a neighbour-joining tree reveal a clear distinction between *bbp* alleles from ST30 and *sdrE* alleles from the remainder of the population (Figure 1). *SdrE* genes from isolates of the same sequence type are identical irrespective of the isolation from disease patients or healthy carriers. This suggests that allelic variation reflects clonal background. The presence of the *sdrE* within these strains was determined in chapter 4 and this sequencing confirms that assay. Although the tree in Figure 1 illustrates the divergence between the *bbp* alleles in ST30 from the remaining *sdrE* alleles, it provides no information regarding the localisation of the divergent sites. Figure 2 compares the polymorphic sites for *sdrE* and *bbp* genes. An excess of the variation is found within the A region of these sequences. This Figure also shows the presence of mosaicism (the non-random distribution of polymorphisms) which is indicative of a history of recombination within this gene. In this alignment we also observe some localised sequence similarities between an *sdrE* allele from strain C730 with the *bbp* allele sequence. This may also be indicative of localised or repeated recombination between the two *sdrE* allelic variants since the small region of identity is flanked by regions of high divergence. Figure 3 clearly illustrates the localisation of variation between the *bbp* and *sdrE* type alleles of the *sdrE* locus within the C terminal

end of the A region. This is of particular interest since this is the biologically active domain in this protein. Variation within the *sdrE* alleles alone appears to be more localised at the N terminal end of the A region. The level of variation within the B repeats is consistent between alleles at this locus and the variation in the repeat R region is attributable to the variation in repeat number between alleles.

Strain	ST	gene size (bp)	number of SD repeats
D363	30	3324	52
B87	25	3414	67
B504	30	3420	68
H513	30	3438	71
C101	30	3438	71
B226	30	3438	71
C730	45	3450	73
C13	22	3486	79
B426	22	3486	79
C521	5	3486	79
D10	5	3486	79

Table 2. *SdrE* gene sizes

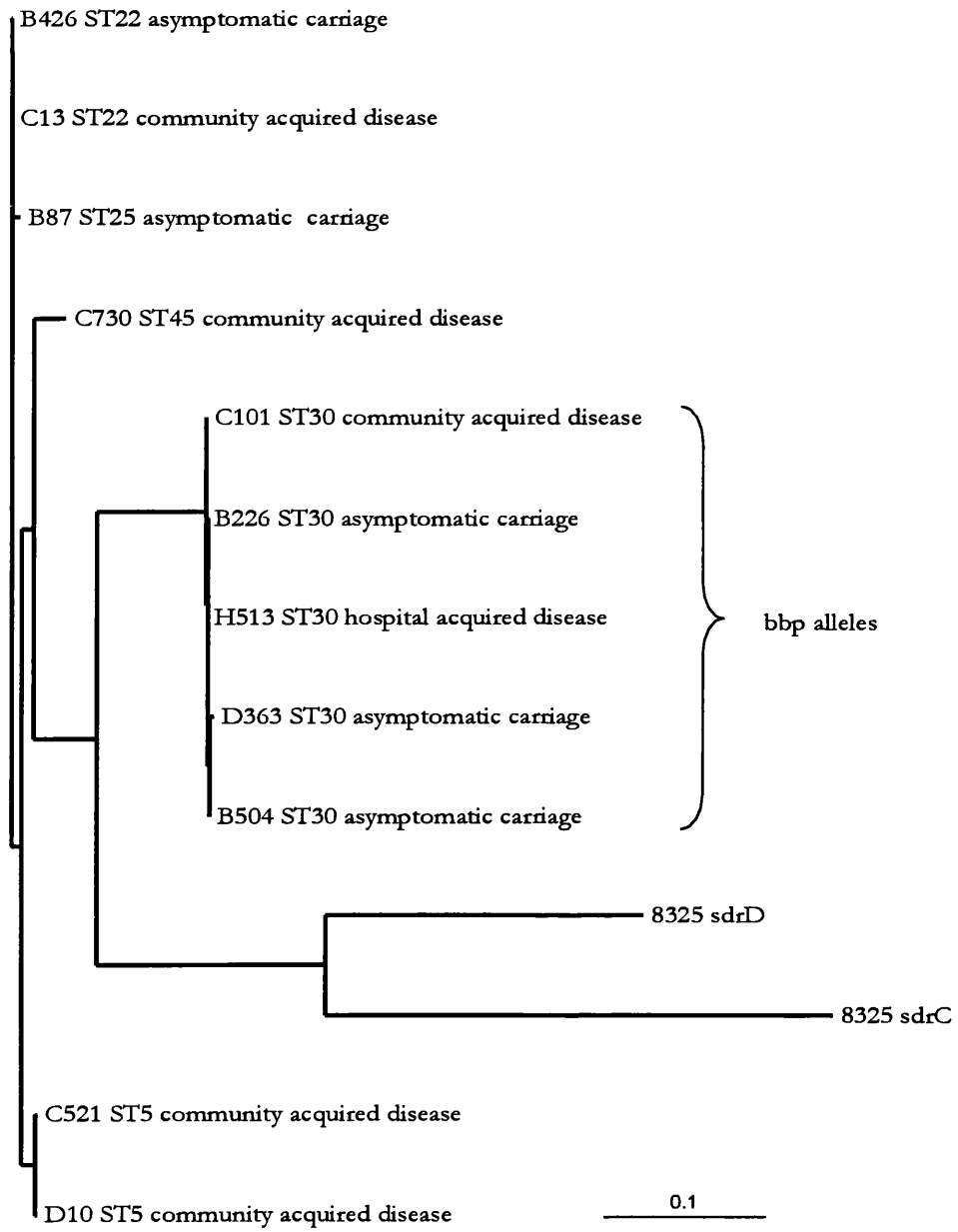


Figure 1. Neighbour-joining tree for translated nucleotide sequence of *sdrE* locus in 11 strains of *S.aureus* with *sdrC* and *sdrD* genes from 8325.

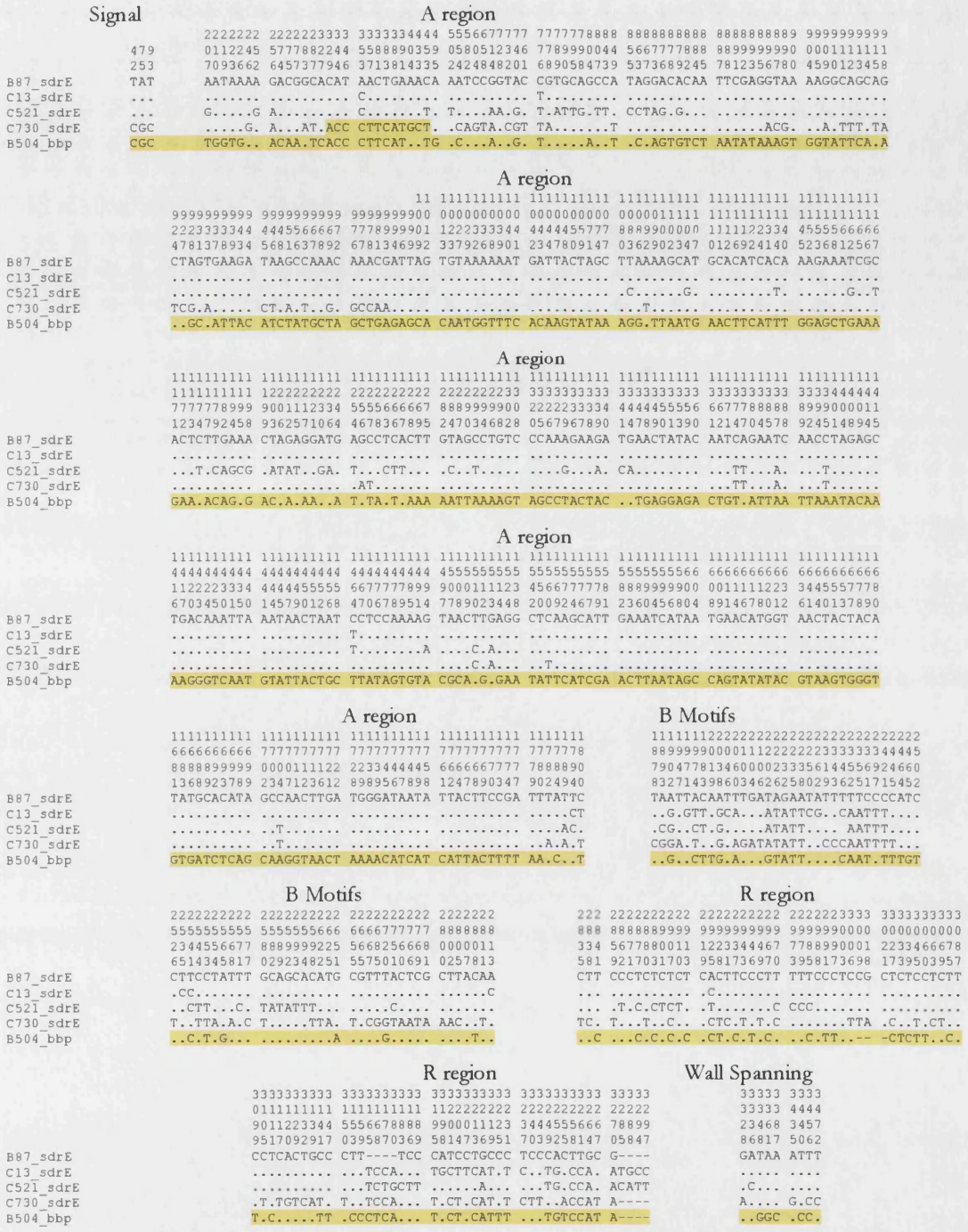


Figure 2. Variation within *sdrE* genes relative to domains. The *bbp* allele sequence is highlighted in yellow as are localised runs of identity between *bbp* and *sdrE* alleles.

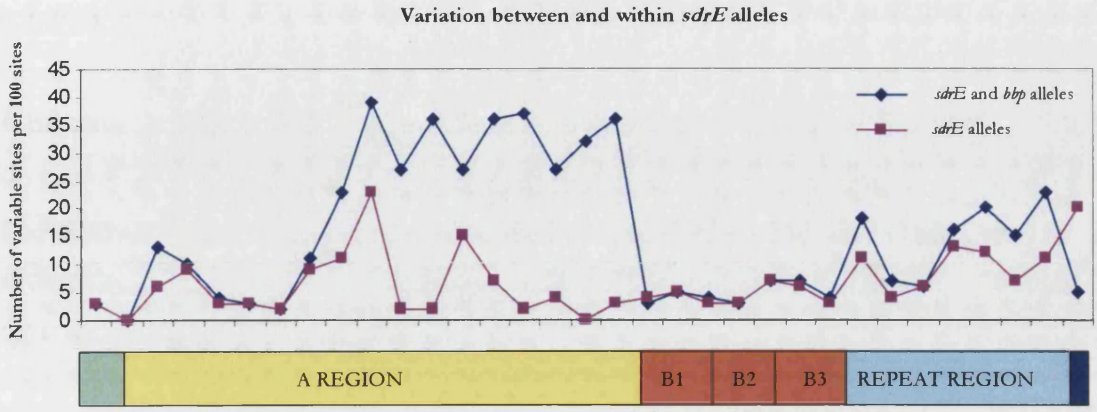


Figure 3. *SdrE* allele variation relative to protein structure.

5.2.2 Variation within the functionally active A region of *sdrE* alleles

It became clear with the reconstruction of a population framework in chapter 3 that highly divergent lineages exist within the natural population of *S. aureus*. ST30 is found in a group of highly divergent lineages based on the sequences of 38 diverse loci. Such genomic divergence may be able to explain the level of variation between *sdrE* alleles and the *bbp* type allele which is found predominantly within lineages ST30 and ST121 and their closely related genotypes. Further characterisation of *sdrE* alleles in the population may provide insight into the origin of the *bbp* allele and its dissemination within unrelated lineages. The localisation of variation within the A region, as shown in Figure 3, is of particular interest since this encodes the ligand binding domain for the sdrE protein. Variation within this domain may have functional consequences and reveal sites of functional importance. In order to further address these issues regarding the evolution and function of *sdrE* alleles, a further 17 strains have been included to represent the diverse lineages of the *S. aureus* population as characterised in chapter 3. The Maximum likelihood tree generated in Figure 4 for the representative population sample reveals the diversity present within the A region of *sdrE* genes. Strains of identical ST group together, and in most cases with 100% identity. There are obvious clades within the *sdrE* alleles, STs representing group 3 cluster together. Group 2 of the population is divided into distinct 2 clades. The divergent lineages which comprise group 1 (ST45 and ST207) of the population lie between these clades. The divergence seen in the *bbp* allelic sequence of ST30 is not relative to other group 1 members and the level of variation seen in the *bbp* allele is greater than would be expected for this lineage. The *bbp* allele from lineages ST101 and ST121 are identical to each other but not identical to the ST30 *bbp* sequence. Lineages ST101 and ST121 are found within different groups of the population and thus this identity has probably arisen as a result of recombination and not descent. The *bbp* sequence from strain C2 representing lineage ST7 is also not identical to that of ST30 *bbp* sequence and would be expected to be more similar to the *bbp* sequence of ST101 than ST121 since ST7 and ST101 lineages are closely related. However, this is not the case and is highly indicative of recent recombination having occurred involving *bbp* alleles.

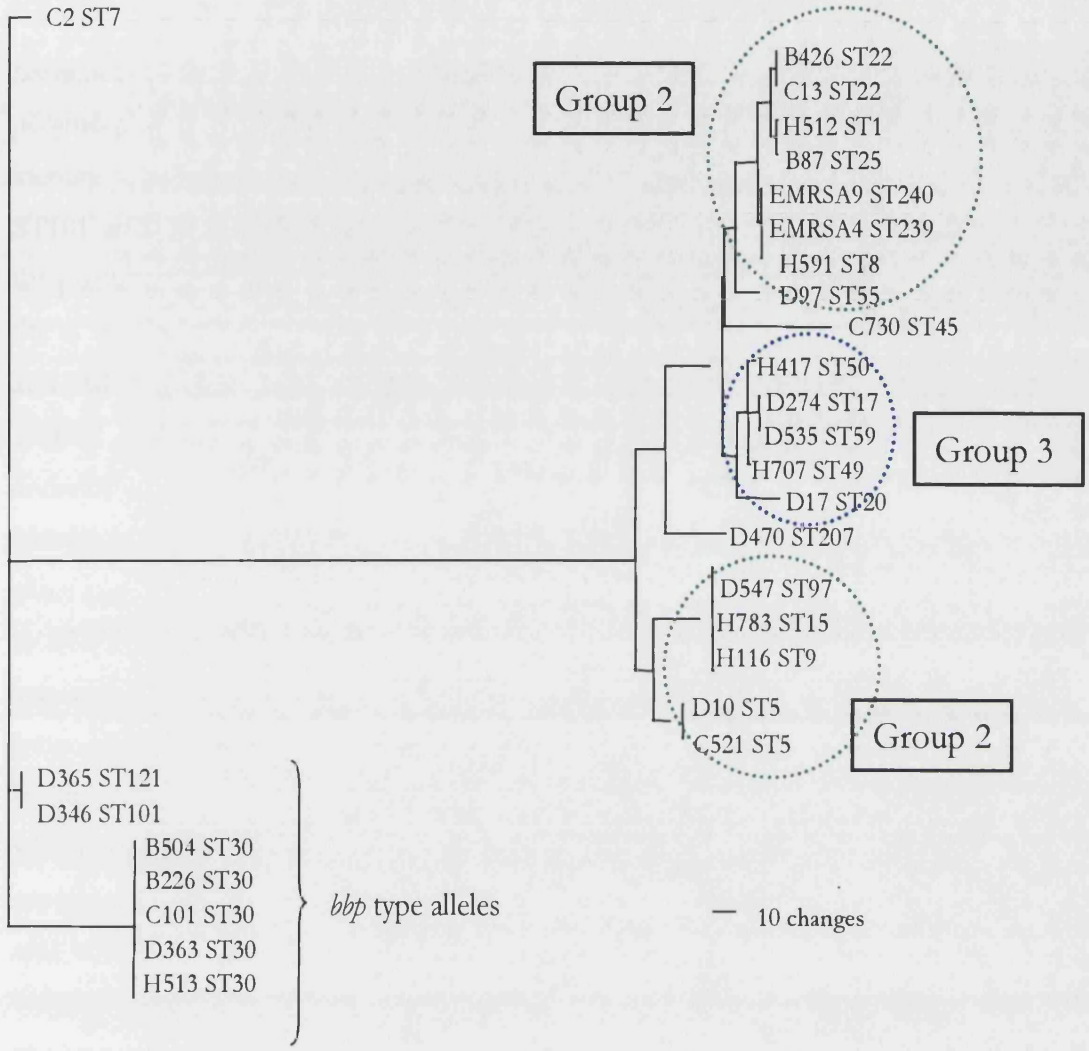


Figure 4. Maximum likelihood tree for nucleotide sequence of *sdrE* A region in 29 strains of *S.aureus*.

5.2.3 Evidence for recombination between and within *sdrE* and *bbp* alleles

Evidence for recombination can be found from the visual inspection of the distribution of polymorphic sites from a set of aligned sequences. The highlighted yellow sequences show identity with *bbp* A region sequences from ST30 in Figure 5. Unrelated lineages ST7, ST101 and ST121 have a high level of identity with ST30 *bbp* sequences, thus validating the positive result from the assay in Chapter 4. This identity with *bbp* begins at base 702 in the A region (base 884 of the gene) before which they resemble other *sdrE* alleles. STs 7 and 101 share a common ancestor in group 2 of the population framework. (see tree chapter 4). This *bbp* fragment is most likely to have been acquired by this common ancestor prior to the diversification of ST101 and ST7. It is interesting however that identity in ST121 of group 3, unrelated to ST7 or ST101, should begin in exactly the same place and have complete identity to ST101. They all also have a small mosaic with identity to STs 5, 15, 97 and 9. This pattern suggests that the sequence was passed between these sequence types rather than in two independent events both originating from ST30. I propose that the most parsimonious pathway is that *bbp* sequence from ST30 was incorporated into the ancestral genome of STs 7 and 101, displacing the original sequence which resembled alleles from ST8 and ST5 respectively, both group 2 lineages. A small mosaic identical to ST5 is still observed before the *bbp* conversion point at base 702. STs 7 and 101 diversified to the state in which we observe them in the population framework today. In a more recent event, a fragment larger than the A region was incorporated into the genome of ST121 from ST101, thus explaining the complete identity between A regions in these unrelated lineages. Localised identity is between unrelated lineages is shown by identical colourings on the population framework shown in Figure 6. The series of events here can be described by parsimony but other potential events are less readily identifiable by eye alone. Needless to say there is extensive mosaicism seen within the A region of this locus and although identity is maintained between very closely related strains and lineages it is evident that there has been some recombination within their ancestors. The impact of recombination will be measured by the implementation of several methods.

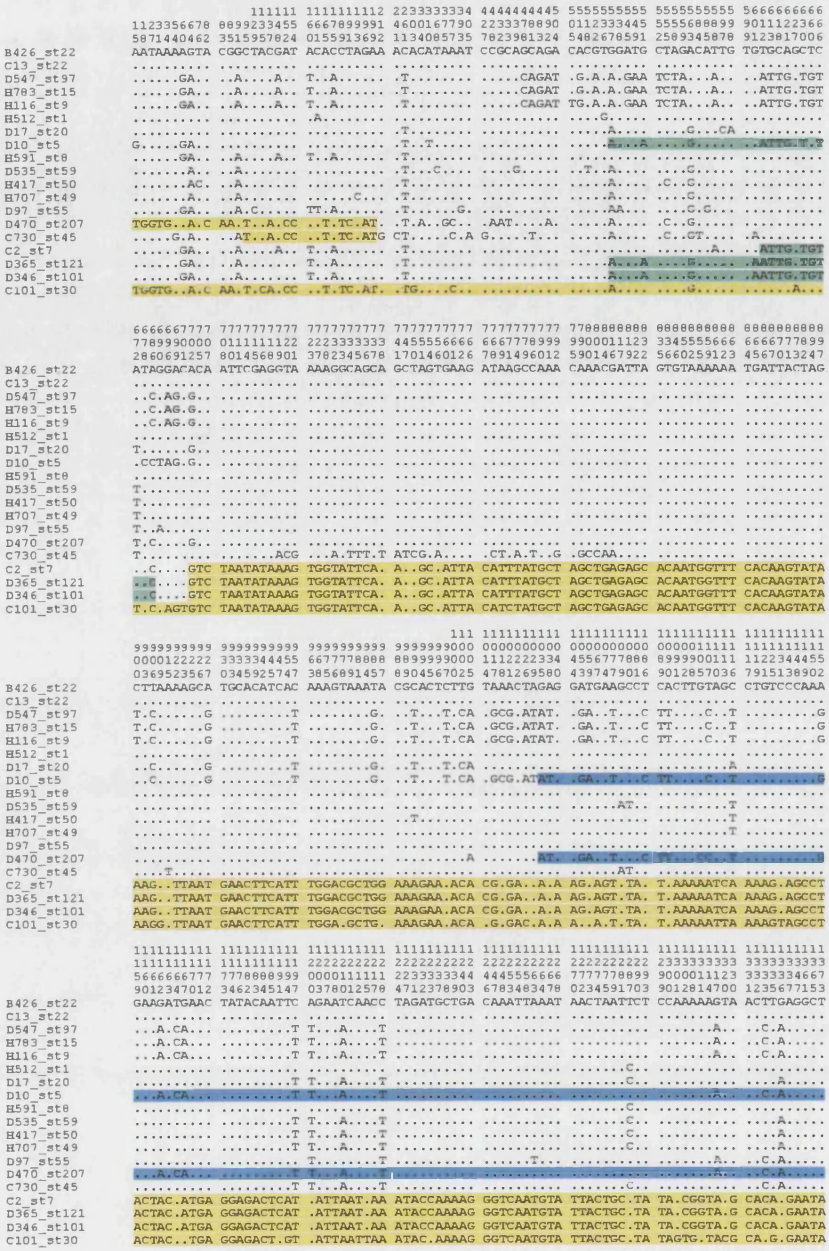


Figure 5. Variable sites within the A region of the *sdrE* gene. Some identical sequences have been removed.

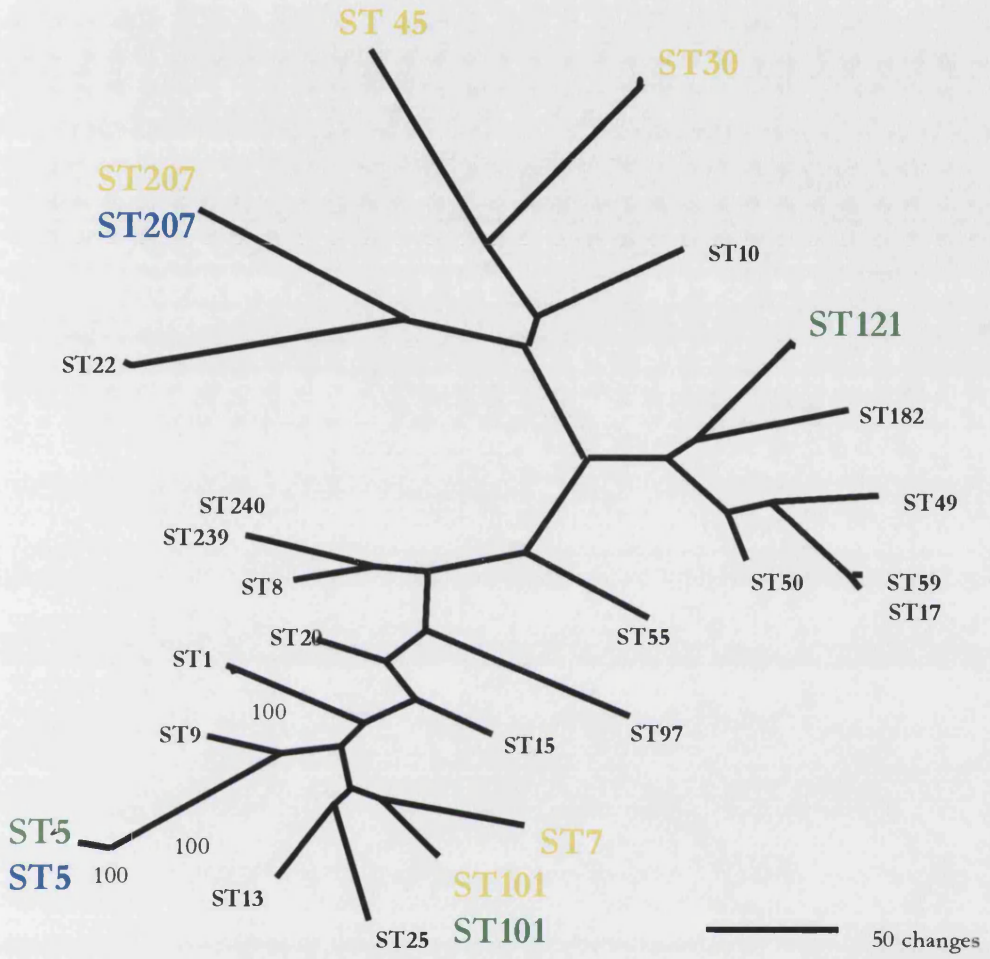


Figure 6. *SdrE* identity within different lineages of *S. aureus*. Identical colours represent localised regions of identity between unrelated lineages.

Tests for recombination within aligned gene sequences have also been implemented to measure the impact of recombination within the *sdrE* A region. The Sawyer's runs test detects significant recombination (p =0.000) from both condensed (SSCF) and uncondensed fragments (SSUF). The largest condensed fragment is 120 bases long, and the largest uncondensed fragment 1614 bases. This result indicates that there has been recombination of different sized fragments and potentially repetitive recombination masking the majority older recombination fragments. Bellerophon looks within multiple sequences for chimeric sequences where both parental sequences must be present within the dataset. Visual inspection of the data suggests that parental sequences are in the data and bellerophon should be able to detect the mosaic sequences. The output tells us there are 138 chimeric sequences within the data. Thus, this technique is limited since recombination has been too extensive for Bellerophon to provide output data The Maximum chi-squared test can be used to calculate the significance of visually identified mosaics. The sequence for lineage ST121 appears to be a hybrid of ST15 (parent 1) and oST30 (parent 2). The mosaic observed here is found to be highly significant (p <0.0001). Another significant mosaic (p <0.0001) is found in the divergent lineage ST207 when compared to ST30 (parent 1) and ST5 (parent 2). DNAsp implements the method of Hudson and Kaplan 1985 to indicate the minimum number of recombination events (R_M) in the history of the sample (note that R_M underestimates the total number of recombination events). For the A region sequences the minimum no of recombination events is R_M =29. This value is much greater than found for any of the ubiquitous genes investigated in chapter 3, although the length of the sequence is at least 3 times longer.

5.2.4 Does the A region of SdrE exist in 3 subdomains?

The A region of model sdr family member, *clfB*, has been reported to exist in subdomains (Perkins *et al.*, 2001). The sdr proteins are structurally similar supporting the existence of such a division of the A region in sdrE. However, the greatest homology with the A region sdrE A region is not with an *S. aureus* sdr protein but with SdrG (or Fbe), a fibrinogen-binding protein of *S. epidermidis* (Hartford *et al.*, 2001). However, homology is limited to the C terminal end of the A region (putative N2 and N3 subdomains).

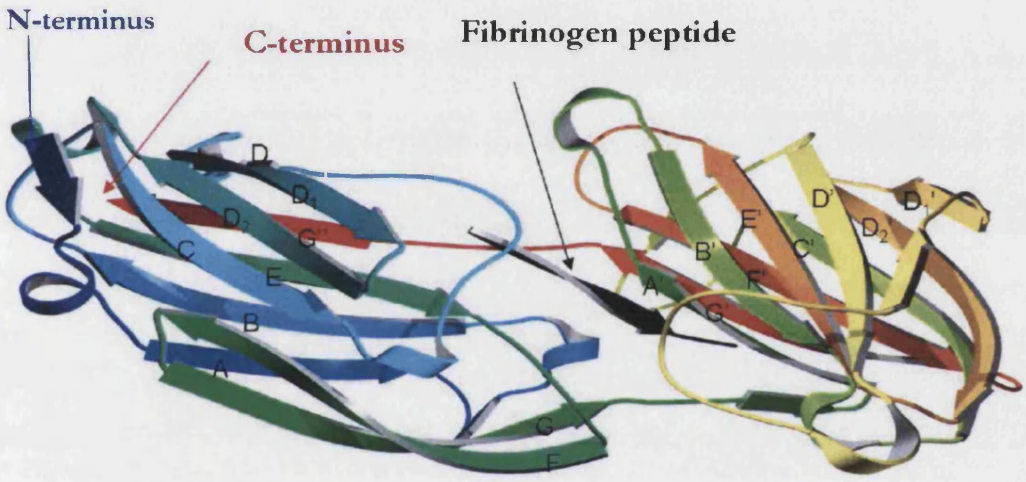


Figure 7. Model of putative N2 and N3 sdrE A region subdomains based upon sdrG (Ponnuraj *et al.*, 2003).

Letters identify antiparallel strands of β-sheets. The start of the putative N2 is located at the N-terminus (blue). The end of the putative N3 is located at the C-terminus (red). The colouring of the sheets and coils indicates the procession of the protein from the N-terminus to the C-terminus in the following order: blue, turquoise, green, yellow, orange, red. The fibrinogen-peptide is also shown to indicate the region in which this ligand is bound to sdrG.

The structure in Figure 7 supports the existence of two subdomains which form two immunoglobulin folds, typical of surface expressed proteins. For sdrG, fibrinogen binding occurs at sites between the immunoglobulin folds. In the case of clfB, binding activity is also localised within the N23 subdomain of the A region. The N1 subdomain of the sdrE A region has little homology with other sdr genes. If the ligand-binding properties are located between the immunoglobulin folds of N2 and N3, then N1 may play a role in the full display of the A region outside of the bacterial cell. The presence of a SLAVA motif in clfB suggest that there may be an element of proteolytic processing of N1. Cleavage by aureolysin was detected between Ser^{197} and Leu^{198} (McAleese *et al.*, 2001). In the sdrE protein we find similar motifs RFAVA and PAAVA suggesting this protein could be processed by aureolysin in a similar manner between $Ala^{197}$and Val^{198} and between $Ala^{199}$and Val^{200}. It is possible that the physical presence of this subdomain may mask the ligand-binding site and so the proteolytic break down of this plays a role in fully exposing the binding site as well as subsequent shedding of the protein.

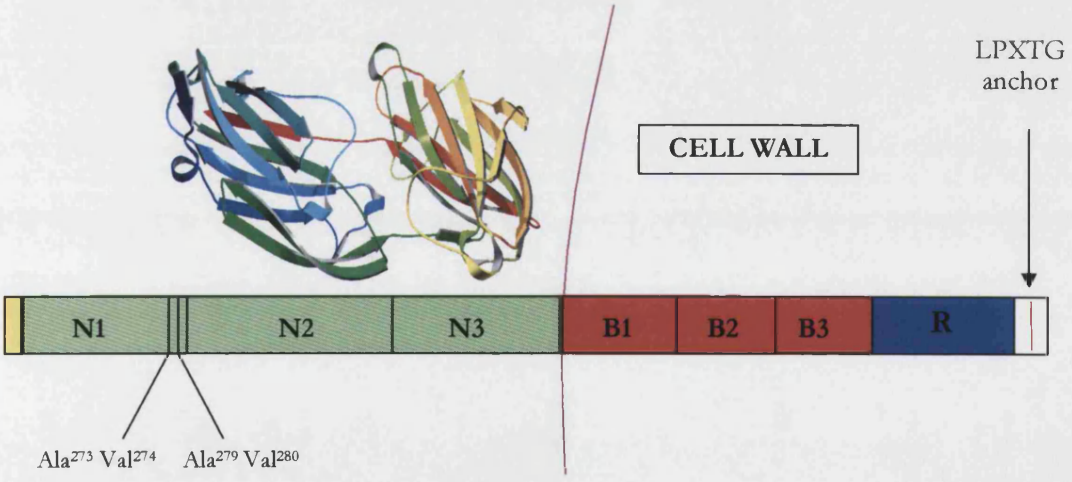


Figure 8. SdrE immunoglobulin folds in the context of the full sdrE protein.

Changes in tree topology across the putative sdrE A region subdomains are shown in Figure 9. In this region of only approximately 1600 bases, the tree topology changes quite dramatically and close associations for some strains, depicted with colours, are dispersed in subsequent domains of the A region. However, we also observe that although the divergence between the sdrE and bbp allelic types appears to increase within the putative N2 and N3 subdomains, there is less diversity within each of these compared to the putative N1 domain. This suggests that the necessity to form these immunoglobulin folds may constrain the accumulation of variation within these subdomains. It is possible that the putative N1 subdomain is less functionally constrained as a result of a less direct role in ligand binding. Alternatively, as an antigenic protein, this domain may be subject to diversifying selection in order to avoid recognition by the host immune system.

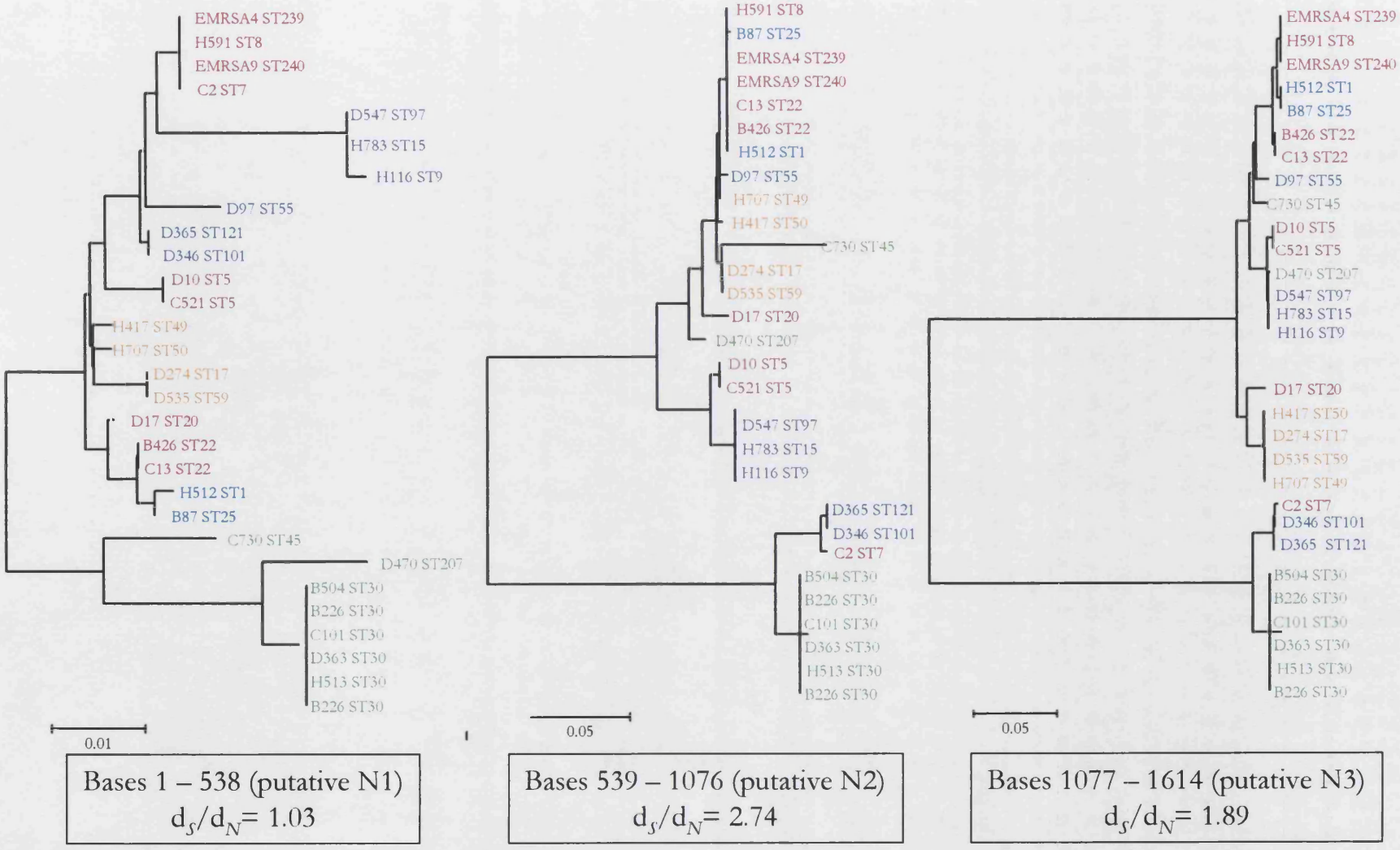


Figure 9. Changes in Neighbour-joining tree topology over the A region of *sdrE* alleles

5.2.5 Evidence for selection at the *sdrE* locus

As *sdrE* is an accessory locus there may be a more relaxed functional constraint compared to ubiquitous essential loci. Surface protein adhesins are exposed directly to the host resulting in diversifying selective pressures acting upon these proteins. The ratio of synonymous to non synonymous substitutions (d_S/d_N) can be used as an indicator of positive or diversifying selection. A value < 1 represents an excess of non synonymous substitutions to synonymous substitutions and suggests positive selective pressure is acting on the protein sequence to change. The calculated d_S/d_N value of 11 complete *sdrE* (*sdrE* and *bbp* alleles) is 2.679. The calculated d_S/d_N value of 30 A region sequences is 2.17. Although these values are > 1 they are low compared to conserved genes and comparable to the ORPHANS examined in chapter 1. This method looks for selection over the entire region of sequence executed. However, selection may be acting upon particular domains within a gene, or upon single sites. Datamonkey, part of the HYPHY package looks for evidence of positive selection at individual sites.. Inputting the sequences of 11 *sdrE* and *bbp* gene sequences, datamonkey identifies 35 negatively selected sites and no positively selected sites. Seven codons are located within the A region, 4 codons in the B repeats and 24 codons in the R region. Such a high number in the repeat region is to be expected due to the conserved SD repeats. Inputting 30 A region sequences, a further 31 negatively selected sites are identified within the A region alone. However, a single positively selected site is also identified within the A region. This is codon 54 of the A region sequence, codon 106 in the complete protein sequence. Interestingly this site is located within the putative N1 subdomain of the A region and is shown in Figure 10.

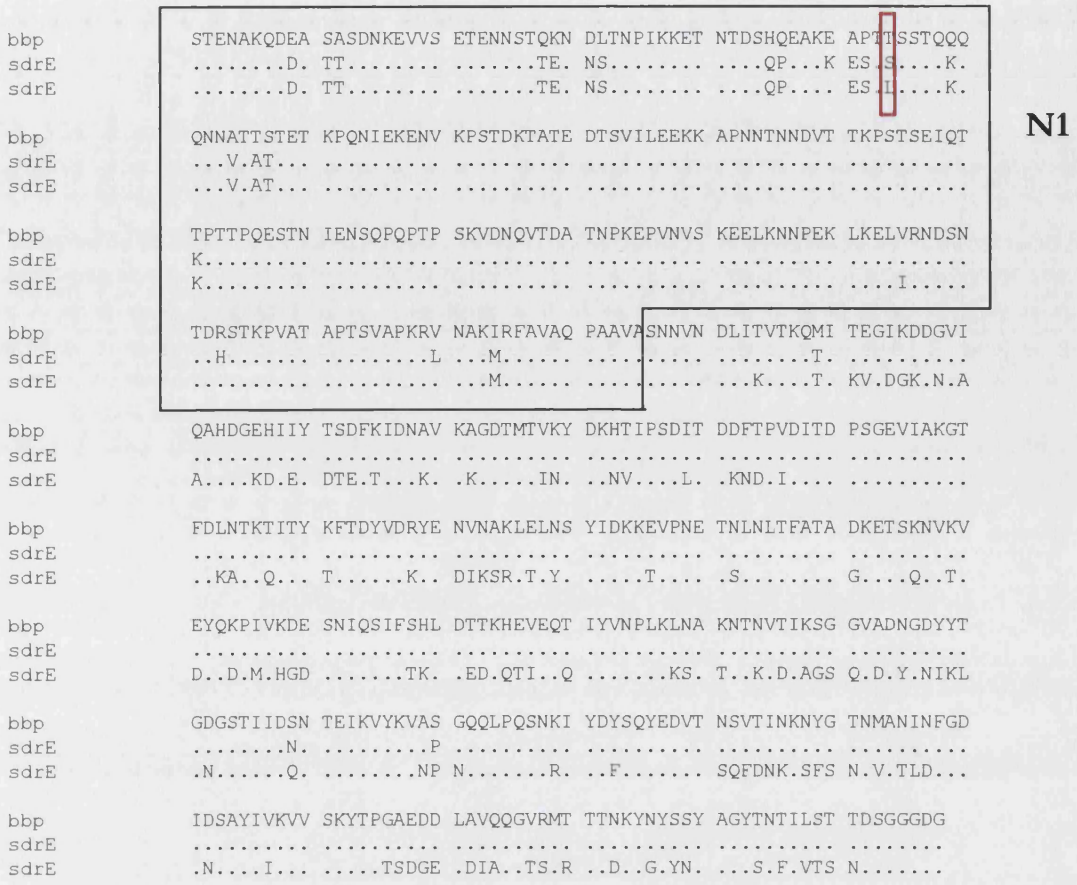


Figure 10. Identification of a positively selected site within the putative N1 subdomain of sdrE.

5.2.6 The functional implication of variation at the *sdrE* locus

The variation seen within this binding region is quite striking, particularly compared to the bbp protein sequence. However, it is unclear whether this variation has any functional implication. The *bbp* allele of *sdrE* has been reported to bind bone-sialoprotein (bsp) , the *sdrE* allele of the *sdrE* locus is reported to play a role in the activation of platelet aggregation (O'Brien *et al*, 2002). The binding of bsp by either *sdrE* or *bbp* alleles has yet to be successfully demonstrated by other laboratories and whether allelic variation is responsible differentiation for this function is unclear. However, the ability of sdrE alleles, including bbp, to activate the aggregation of platelets is as yet unreported and is tested here. Two strains for which complete gene sequence had been generated, C101 and C730, were used for functional analysis by cloning and expression of the gene in *Lactococcus lactis.* The protein sequence of the A regions are shown in Figure 11. Strain C101 expresses the bbp type allele and C730 expresses a divergent sdrE protein with some unique residues.. These will be tested for functionality along with the sdrE from laboratory strain Newman. These strains represent diverse sdrE proteins. A further strain C767, known to have both *sdrE* and *bbp* alleles, was also used with the secondary aim of sequencing the cloned *sdrE* or *bbp* (or both) without the hurdle of a mixed PCR product. Images representing the results obtained in the cloning and transformation of *Lactococcus lactis* are shown in Figures 12 -16. Detailed methods for cloning, expression and aggregation experiments can be found in Chapter 2 from page 46.

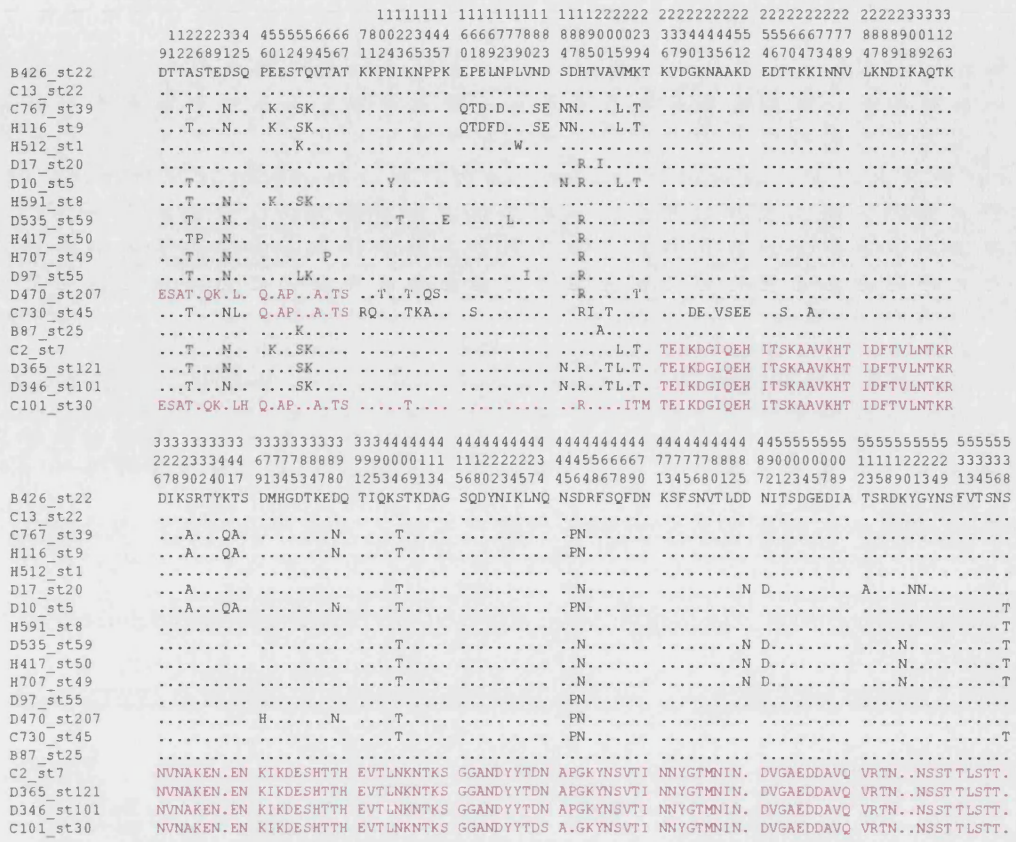


Figure 11. SdrE A region protein alignment. Bbp unique sequence is highlighted in red, as is homology with bbp in sdrE sequences.

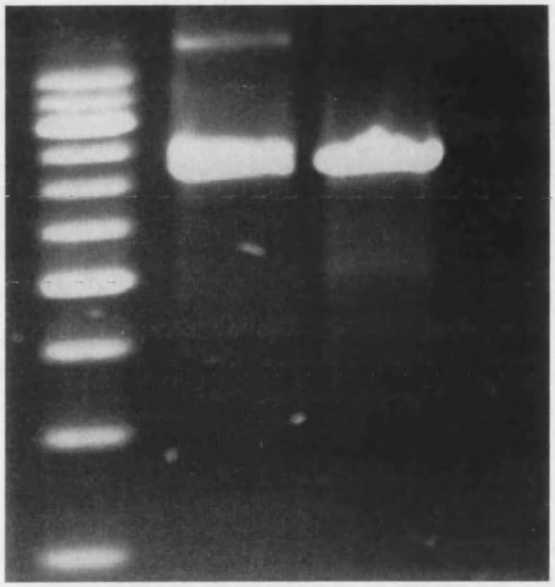


Figure 12. *SdrE* gene amplicons run on 0.8% agarose TAE gel.

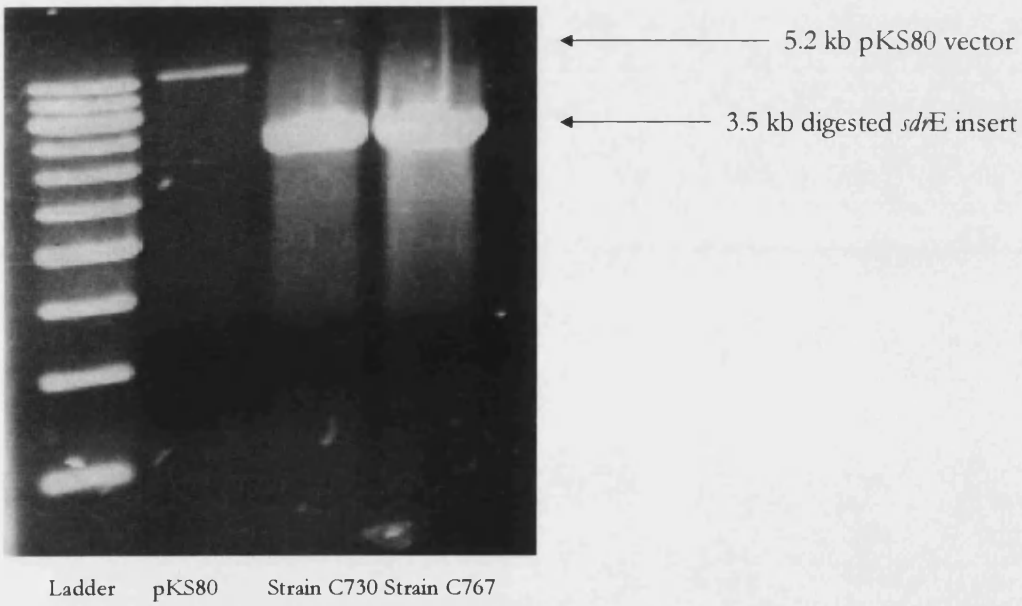
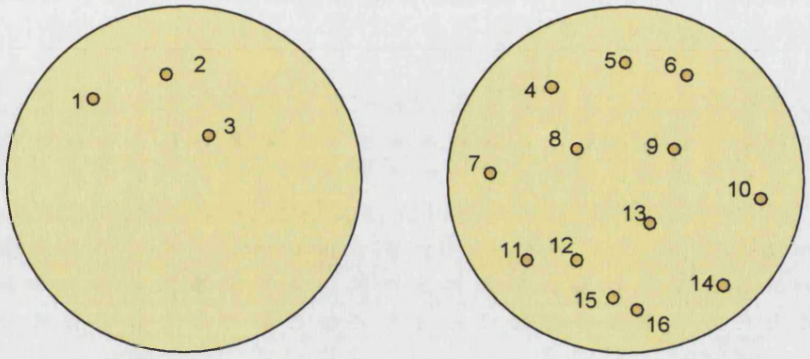


Figure 13. Comparitive 0.8% agarose TAE gel of pkS80 vector and digested *sdrE* gene amplicons.



L. Lactis colonies
pkS80 vector with *S. aureus* strain C767

L. Lactis colonies
pkS80 vector with *S. aureus* strain C730

Figure 14. Transformed *L.lactis* colonies on GM17 agar with erythromycin.

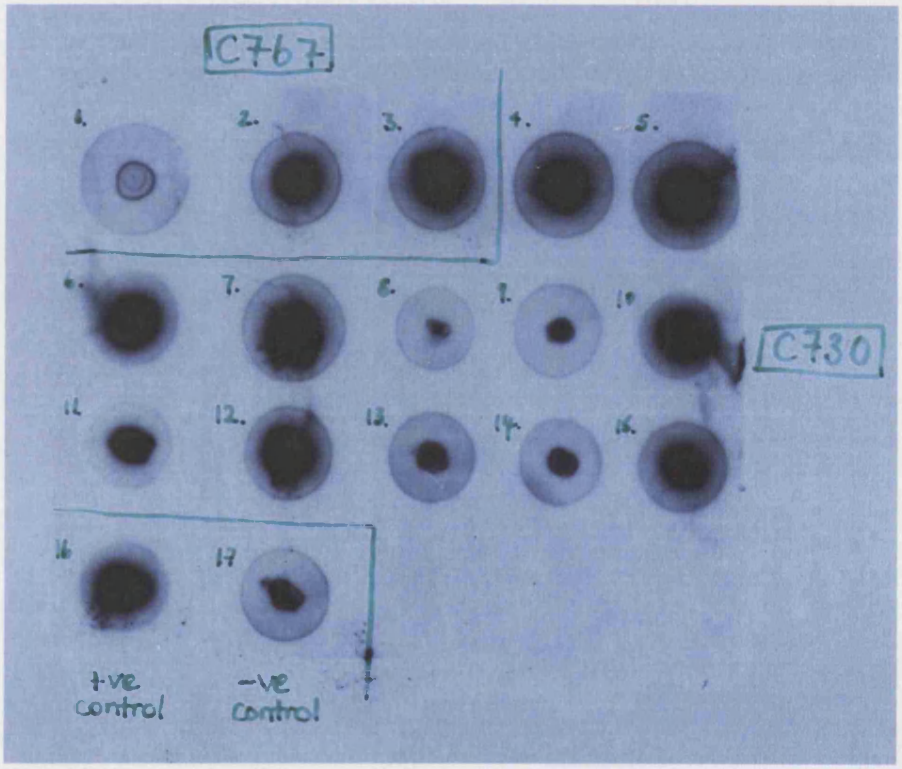
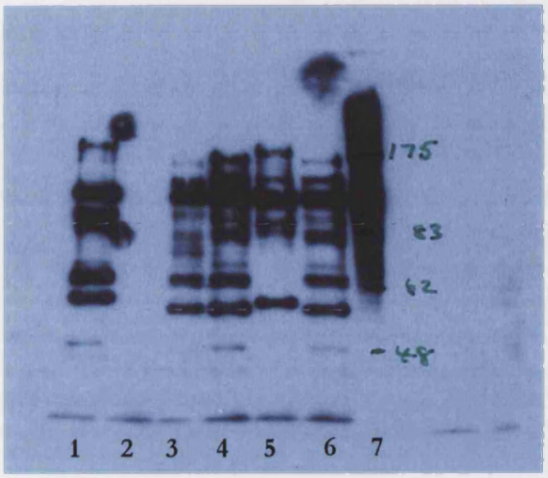


Figure 15. SdrE expression screening of transformed colonies.

Using controls for comparison:

L.lactis colonies 2 and 3 are expressing sdrE from *S. aureus* strain C767

L.lactis colonies 4, 5, 6, 7, 10, 12 and 15 are expressing sdrE from *S. aureus* strain C730



1 - positive control
L. lactis with sdrE from *S. aureus* Newman (ref)

2 - negative control
L. lactis wild type

3 - *L. lactis* colony # 5 (see figure 4)

4 - *L. lactis* colony # 7(see figure 4)

5 - *L. lactis* colony # 2 (see figure 4)

6 - *L. lactis* colony # 3 (see figure 4)

7 - Ladder

Figure 16. Western immunoblot of sdrE proteins.

A protein of approximately 175 KDa is expected in lanes 1, 3 – 6. Surface proteins have been successfully expressed and subsequently removed from *L. lactis*. The presence of multiple bands in these lanes is a result of further enzymatic degradation of the proteins subsequent to their release from the bacterial cell wall.

CHAPTER FIVE: CHARACTERISATION AND FUNCTIONAL IMPLICATIONS OF VARIATION AT THE SDRE LOCUS

The successful expression of sdrE proteins (including bbp) in *Lactococcus lactis* enabled the controlled testing of the effect of diverse protein sequences on the ability of *S. aureus* to activate the aggregation of platelets in human plasma.

The ability of *Lactococcus lactis* MG1363 cells, expressing sdrE, to activate platelet aggregation

Wild type *S. aureus* has the fastest activation time. *S. aureus* expresses not only sdrE but a plethora of surface proteins with the ability to activate the aggregation of platelets. All *sdrE* alleles, including bbp, which have subsequently been expressed in *L. lactis* have shown the ability to activate the aggregation of platelets (Figure 17). Wild type *S. aureus* is capable of the greatest percentage aggregation of platelets. Wild type *Lactococcus lactis* is unable to activate the aggregation of platelets (Figure 18). There is variation in the percentage aggregation of 19%, and up to 9% between different platelet donors. There is also some variation in the time to activation of 3.8 minutes, and up to 1.7 minutes variation between different platelet donors. The significance of these differences in the time to activation and the percentage aggregation, in a clinical setting is unknown. However, the variation found between different donors suggests that other factors regarding host specificity and measurement error also contribute to the differences observed. The sites involved in the activation of the aggregation of platelets are unknown. However, the variation observed in the bbp allele has not affected this particular function of the sdrE protein. We can speculate that the localisation of variation in the N23 subdomains of the A region result in some conformational changes of the immunoglobulin folds whilst having no effect upon this particular function of the sdrE protein.

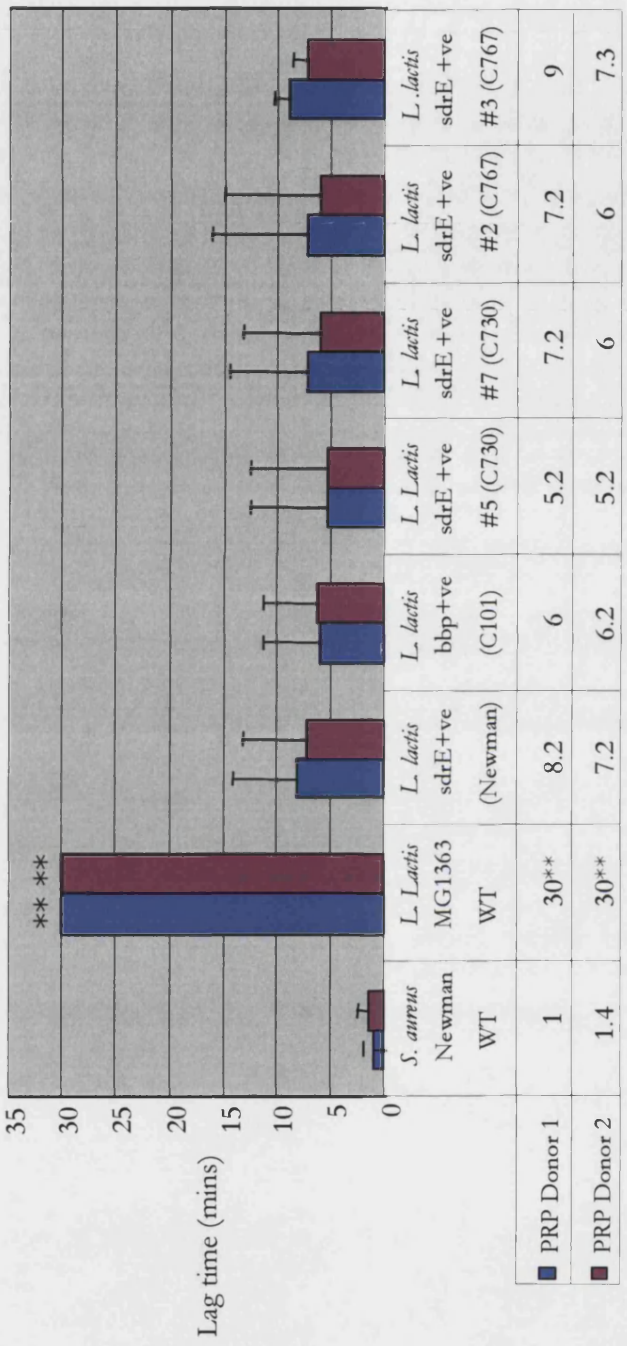


Figure 17. Time to activation of the aggregation of platelets

** indicates that aggregation did not occur after 30 minutes incubation

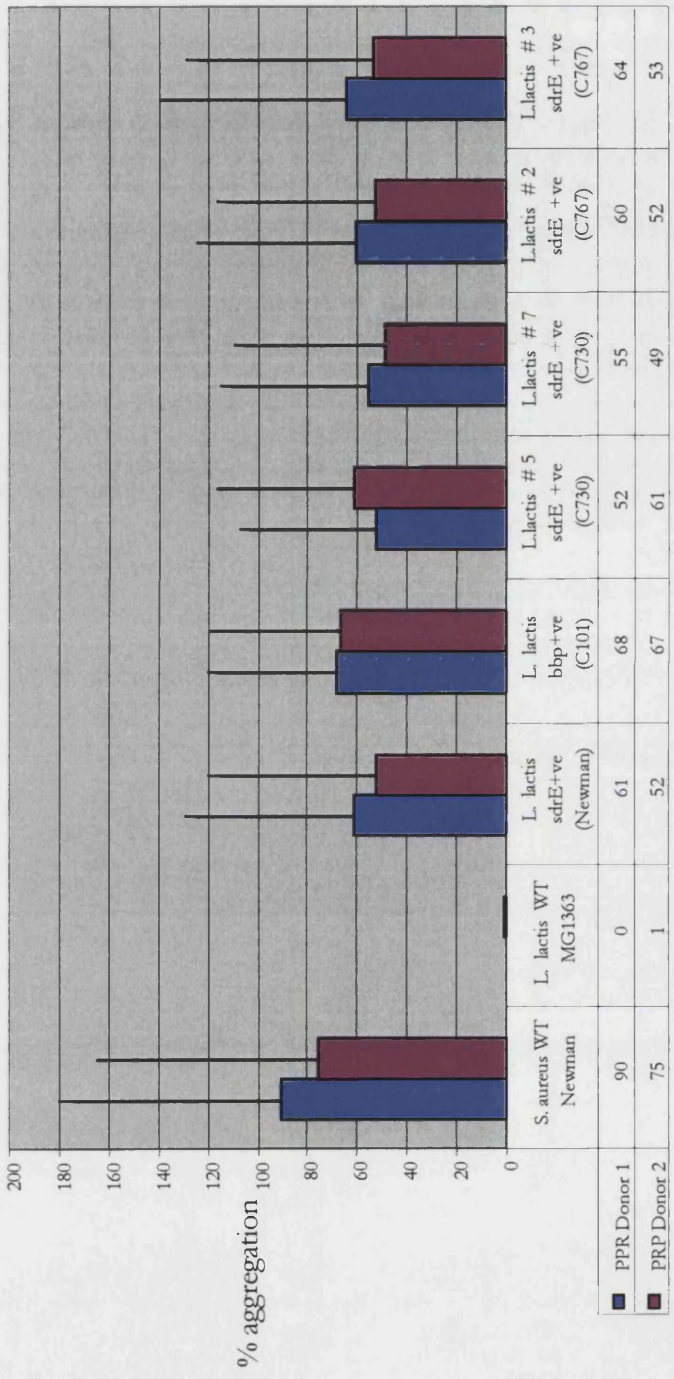


Figure 18. Percentage aggregation of platelets

Several strains within CC30/39 were found to have two copies of this locus representing both *sdrE* and *bbp* alleles. The allelic type from strain C767 expressed by *Lactococcus lactis* was determined using the original *sdrE/bbp* assay from Chapter 4. It was *sdrE* positive and *bbp* negative. This amplicon was then sequenced to determine homology to any other *sdrE* alleles within the population. The A region sequence was obtained as had been done for all other isolates. When included in the phylogeny generated for *sdrE* alleles the sequence of the *sdrE* A region found in strain C767 (ST39) was identical to that in STs 97, 15 and 9 (Figure 19).

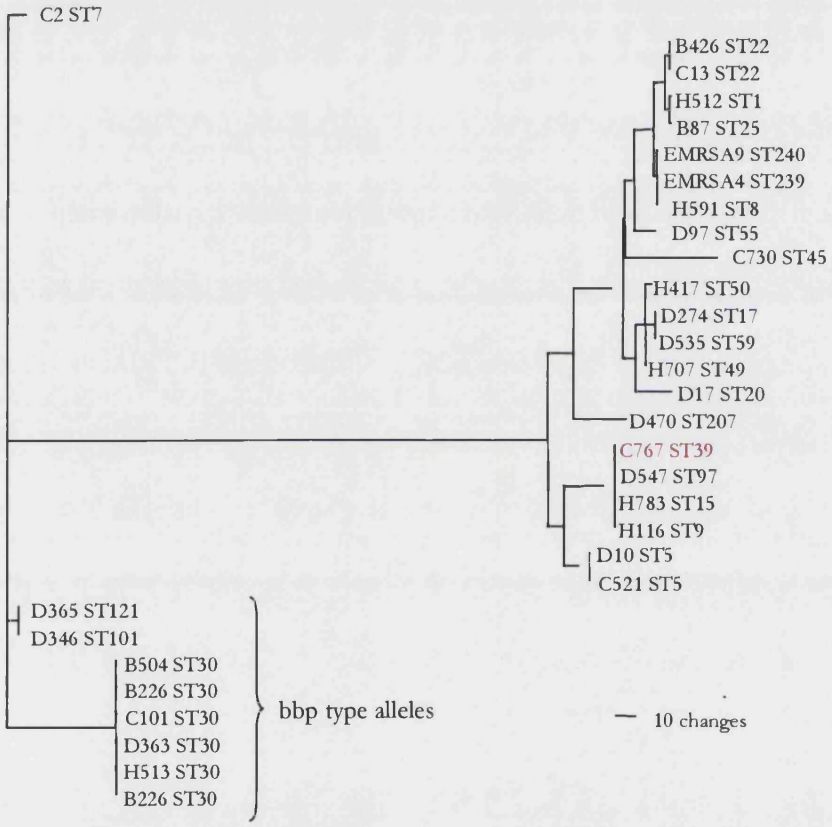


Figure 19. Maximum likelihood tree for nucleotide sequence of *sdrE* A region sequences including *sdrE* from ST39.

5.3 DISCUSSION

In this Chapter we have shown the localisation of variation between *sdrE* and *bbp* allelic types of the *sdrE* locus within the functionally active A region of the encoded surface protein. We find no evidence for a difference in the ability of *sdrE* allelic variants to activate the aggregation of platelets. Platelets are considered to play a primary role in the early steps of staphylococcal adhesion to damaged endothelium and to spur on the growth of the infective vegetation (Sullam *et al.*, 1996; Yeaman *et al.*, 1996). In this way this particular bacterial function has great implications for the development of infective endocarditis. The ability of allelic variants to activate the aggregation of platelets and the positive association of this locus with isolates from disease suggests that this function plays an important role in Staphylococcal pathogenesis. However, sdrE is not the only *S. aureus* surface protein implicated with a role in the activation of platelet aggregation. Compared to the clumping factor proteins sdrE has a relatively long lag time to the activation of platelet aggregation (O'Brien *et al.*, 2002a) and these other proteins would presumably compensate for a loss of the *sdrE* locus in some lineages. However, surface proteins typically have more than one function and it is possible that there are further uncharacterised roles which contribute to the association of this locus with disease isolates. Repeated assays, in two laboratories, have been unable demonstrate binding of recombinant bbp and bone-sialoprotein (bsp) and thus reproduce previous reports the data (Tung *et al.*, 2000). Bsp binding assays for sdrE were also unsuccessful (Mary Meehan and Timothy Foster, personal communication) and so whether there are allelic differences regarding this particular function remains unclear.

The implications of such accumulation of variation by recombination in accessory loci may be complex. Accessory genes may serve an adaptive function and the loss of such loci, as we have seen for the *sdrE* locus may only impact the invasive potential and not the reproductive potential of the individual. However, the presence of other surface proteins which can act with in a compensatory way greatly reduces the impact of the loss of this locus for this particular function. This functional redundancy may restrain the functional constraint enabling the diversification increasing chances of adaption. In Chapter 3 we

observed no significant association between recombination and functional constraint. This was largely accountable by the presence of synonymous recombinational replacements. We observe the presence nonsynonymous mosaics in sdrE. In this case there is likely to be a link between functional constraint and nonsynonymous recombination. In such a locus the implications of variation at a single site may have limited implications and in fact represent extreme relaxed functional constraint as a result of functional redundancy. The significance of the accumulation of nonsynonymous variation is adaptation. In order to adapt to new environments it is likely that new traits and phenotypes are required and these are more likely to be a result of nonsynonymous variation. The tandemly arrayed *sdrC*, *sdrD* and *sdrE* genes presumably arose by ancient duplication and have subsequently diversified. The sdrC and sdrD proteins do not have the ability to activate the aggregation of platelets and the long-term accumulation of nonsynonymous variation in these loci has presumably resulted in their functional differences. The presence of a positively selected site in the N1 subdomain of the A region suggests that this may be an antigenic region of the protein. However, this is a single site and could equally represent an extremely relaxed functional constraint since it is unlikely that if this were the case that only a single site would be involved for diversifying selection. The role of the N1 subdomain in this protein is unknown. In clfB the enhanced binding affinity of the N23 recombinant protein in the absence of N1 suggests that this may be proteolytically cleaved after exposure from the cell wall. The N terminal may accumulate variation as a result of diversifying selection to avoid recognition by the host immune system but without altering the ligand binding capabilities and affinities of the C terminal end. Many negatively selected sites were detected within this domain representing a strong stabilising selection which may be acting on residues involved in binding and in preserving the structural conformation of the immunoglobulin folds and thus the integrity of the binding sites.

The initial PCR assay in Chapter 4 revealed the presence of both *bbp* alleles and *sdrE* alleles within some strains of the CC30/39 clonal complex. One possibility for the presence of both alleles would be the duplication of *sdrE* to result in the *bbp* form. Closer inspection of the relationships and distribution of these isolates with both alleles (2 loci) within the clonal complex reveals that these duplication events must have occurred at least three

times within ST30, ST39 and ST36. The remainder of the isolates within this complex where the bbp allelic form predominates would have had to subsequently lose the original form of the *sdrE* locus to explain the distribution that is now observed. The isolation, cloning and subsequent sequencing of the *sdrE* allele from strain C767 appears to quash this hypothesis. The *sdrE* gene was found to be identical to that found in more closely related strains of group 2. The resolution provided by the population framework in Chapter 3 confirms that such identity is not a result of descent but by recombination or gene transfer. In light of these findings the most parsimonious explanation would be that a copy of the *sdrE* gene from ST15, ST97 or ST9 has been incorporated into the genome of this strain with both forms. In further support of this we can consider the fate of duplicated genes. We observe extensive localised variation between *sdrE* and *bbp* allelic types. Had a duplication event occurred would we expect to see this particular pattern of variation? It seems unlikely that the diversification of a duplicated gene would result in such extensive localised variation whilst maintaining conserved gene and domain sizes and regions of much greater homology with the original *sdrE* alleles. The length of the branch between *sdrE* and *bbp* in ST30 appears much longer than is observed for any of the single genes that have been analysed in this thesis. It is also much longer than is found for other divergent lineages of group 1 of the population. The presence of *bbp* in CC121 and lineages ST7 and ST101 also generates atypical branch lengths for this genotype within its population group 3 (Figure 19). It seems feasible that the variation observed within the *bbp* has arisen as a result of homologous recombination within the ancestor of one of these lineages. Identification of the original recombinant ancestor is difficult. However, this can be speculated with the inclusion of two assumptions: Firstly, that the incorporation of a diverse *bbp* type allele into the *S. aureus* population, potentially from a closely related species, occurred only once and the distribution within the *S. aureus* population is a result of intraspecific recombination. Secondly, the long internal branches represent older lineages of the population. This was speculated in Chapter 3. Based on these assumptions the most parsimonious evolutionary history of the *bbp* type allele based upon tree topology, the population framework and the distribution of polymorphism is as follows:

The *bbp* allele arose within the common ancestor of the clonal complex cc30/39 replacing the original *sdrE* sequence at that locus. This is subsequently observable in the distribution of the *bbp* type allele within this clonal complex. Homologous recombination then occurred with the incorporation of a large but incomplete fragment of the A region into the ancestor of lineages ST7 and ST101. This ancestral sequence subsequently diversified to generate the two lineages of ST7 and ST101 observable in the population framework. ST101 then became the donor for recombination with ST121 *sdrE* sequence. However in this event a larger fragment including the entire recombinant A region was incorporated into ST121. This is presumably the most recent of all the recombination events since complete identity is maintained in the sequenced region of the locus between ST101 and ST121. The redundancy, meaning similarity of sequence, structure and functionality, within the sdr family and other MSCRAMMS suggests multiple ancient duplication events of these genes to create the array of surface proteins that we observe expressed in *S. aureus*. It is speculated that gene duplication events are more likely to produce novel gene function than the accumulation of point mutation. Teamed with the evidence for extensive recombination within such proteins and the existence of divergent alleles supports the idea that duplicated genes can evolve rapidly and potentially acquire novel function supported by the functional redundancy.

The presence of the directly-repeated sequences within the R regions of these tandemly arrayed genes suggests that slipped-strand mispairing (SSM) is a likely mechanism for the existence of varying numbers of repeats within this gene. The peculiar tertiary structure of repetitive DNA allows mismatching of neighbouring repeats and depending upon the strand orientation, repeats can be inserted or deleted during replication (Coggins & O'Prey, 1989). The analysis of variable numbers of tandem repeat (VNTR) has included the MSCRAMMs *sdrC*, *sdrD*, *sdrE* and *clfA* have been included in an PCR-based multilocus VNTR (MVLA) scheme (Sabat *et al.*, 2003). The authors report discriminatory power comparable to pulsed-field gel electrophoresis (PFGE). However, the power of this scheme may be limited to more recent evolutionary relationships since we observe variation in the number of repeats of sdrE between strains which are identical by MLST (Table 2).

The MSCRAMMS are considered to be possible vaccine candidates. The strength of the immune response by sdrE is unreported, although sdrE antibodies have been raised in rabbits (McAleese *et al.*, 2001; O'Brien *et al.*, 2002). The identification of only one positively selected site within the surface exposed A region suggests that the immune response may be weak for this particular protein. However, the clumping factor proteins which have a much faster lag time to the activation of platelet aggregation may be better vaccine candidates. However, vaccine design may still be problematic as a result of redundancy and the extensive variation we have observed in a representative surface protein, sdrE. Recombination within these loci could render a vaccine ineffective unless the full extent of variation within the population is characterised as has been done here. Whether other sdr family members reveal a similar pattern of variation resulting from extensive recombination will be of interest and insights into this will be provided as we investigate the impact close proximity to genes encoding surface exposed proteins has on neighbouring genes in Chapter 6.

CHAPTER SIX

EVALUATION OF MLST 'CORE' AND SDR 'ACCESSORY' NEIGHBOURS

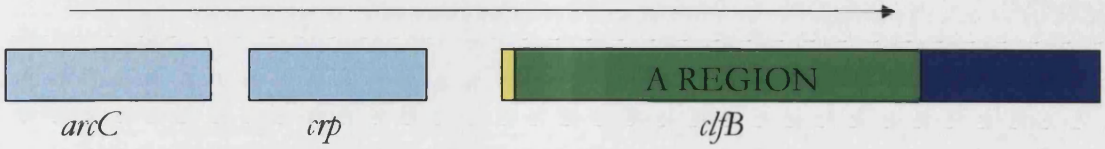
6.1 INTRODUCTION

The aim of this Chapter is to characterise the functional domain of SDR family member clumping factor B (*clfB*) and determine whether proximity downstream of this locus can explain the low level of congruence observed for the MLST housekeeping loci *arcC*. The *arcC* gene lies approximately 2.5 kb downstream of *clfB*. *arcC* is the least congruent of the MLST genes when scored for congruence with each other (Feil *et al.*, 2003) and scores poorly by the SH test when compared to a consensus phylogeny of 37 genes (Table 8, Chapter 3). The function of the clfB protein suggests that this may be a hypervariable locus with evidence for recombination, as has been seen for sdrE (Chapter 5). The *clfB* locus is ubiquitous in *S. aureus* isolates and several ligands for the encoded proteins have been reported, including adherence to cytokeratin-10 which is found in human desquamated nasal epithelial cells (O'Brien *et al.*, 2002). Such a function implies that this particular locus is an important determinant of *S. aureus* nasal colonisation and carriage.

Strain	ST	History	Resistance Profile
H512	1	Hospital -acquired disease	MSSA
H466	5	Hospital -acquired disease	MSSA
C2	7	Community-acquired disease	MSSA
H591	8	Hospital -acquired disease	MSSA
H116	9	Hospital -acquired disease	MSSA
H19	10	Hospital -acquired disease	MSSA
H402	13	Hospital -acquired disease	MSSA
H783	15	Hospital -acquired disease	MSSA
D274	17	Asymptomatic carriage	MSSA
D17	20	Asymptomatic carriage	MSSA
C640	22	Community-acquired disease	MSSA
C720	22	Community-acquired disease	MRSA
C437	25	Community-acquired disease	MSSA
C101	30	Community-acquired disease	MSSA
H325	36	Hospital -acquired disease	MRSA
H295	45	Hospital -acquired disease	MSSA
H707	49	Hospital -acquired disease	MSSA
H417	50	Hospital -acquired disease	MSSA
D97	55	Asymptomatic carriage	MSSA
D535	59	Asymptomatic carriage	MSSA
D547	97	Asymptomatic carriage	MSSA
D456	101	Asymptomatic carriage	MSSA
H560	121	Hospital -acquired disease	MSSA
D365	121	Asymptomatic carriage	MSSA
D22	182	Asymptomatic carriage	MSSA
D470	207	Asymptomatic carriage	MSSA
EMRSA4*	239	EMRSA type strain	MRSA
EMRSA9*	240	EMRSA type strain	MRSA

Table 1. Bacterial strains representing the diversity of the natural population.

Sequencing of the *arcC*/*clfB* region provided a 3,480 kb sequence for analysis in 28 strains representing diverse genotypes. All PCR and sequencing primers can be found within Appendix D1. The sequence is comprised of *arcC* from the start of the MLST allele, the *crp* gene, the A region of the *clfB* gene and the two intergenic regions as shown in Figure 1.



gene	size	base start	start sequence	end sequence
arcC	771	1	ttattaatcc	ttcacatgtag
intergenic 1	95	772	cactttatc	ggagtttaa
crp	708	867	cttatgacag	gtatgtgtttaa
intergenic 2	346	1575	tatacaatgt	tggagtaata
clfB	1560	1920	tttttgaaaaaaa	tacagta
region size	3480			

Table 2. Sequencing strategy and break down of coding and intergenic regions within the 3,480 kb sequence.

6.2 RESULTS

6.2.1 The distribution of variation from *arcC* through *clfB*

The distribution of variation in the region of *arcC*, *crp* and *clfB* regions is illustrated below in Figure 2. The number of polymorphic sites within 100 base adjacent windows of sequence is shown in the y-axis along the 3,480 kb region represented along the x-axis. The highest number of variable sites within a 100 base window for both *arcC* and *crp* is 7 compared to 34 within the *clfB* sequence. There appears to be a gradual increase in variation across the region however, there are small regions of conservation interspersed between regions of variation which may relate to differences in regions of functional constraint even within the same locus. Even *clfB* which contains the highest levels of variation has a region of high sequence conservation within the first few hundred bases of the gene representing the conserved signal sequence. There is also region of conservation between two high peaks of variation within the *clfB* A region which may represent differences in functional constraint within the A region subdomains, as has been previously suggested for the *sdrE* gene in Chapter 5.

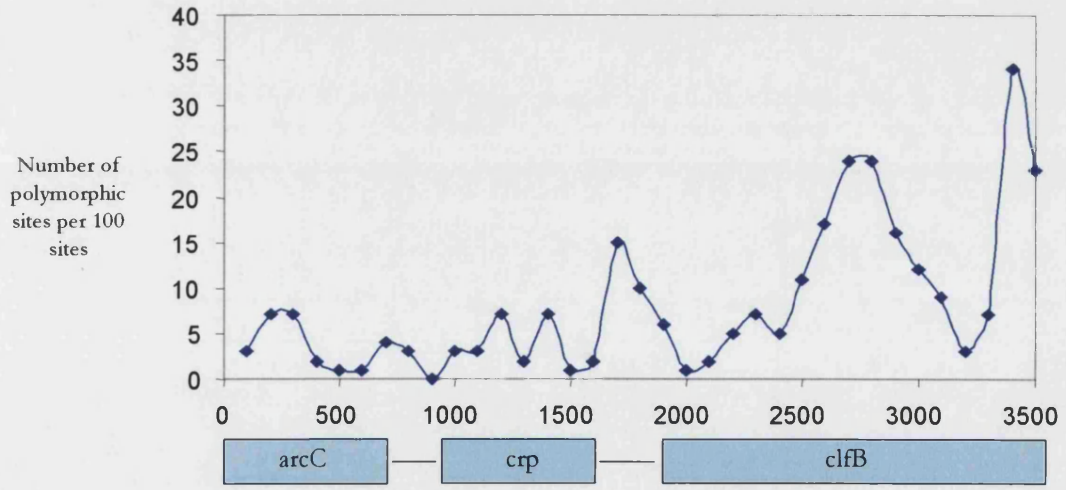


Figure 2. Distribution of polymorphic sites across the region of *arcC*, *crp* and *clfB*.
The number of polymorphic sites within 100 base windows is recorded for the 3.5Kb sequence and shown in relation to the gene arrangement.

The coloured sequences in Figure 3 illustrate examples of identity between the sequences for the *arcC*, *crp* and *clfB* region of strains from different STs. Again, we observe that the variation in this region lies predominantly within the sequence for *clfB*. The *clfB* region is highly mosaic and for the *clfB* sequences alone Bellerophon indicates the presence of 138 chimeric sequences. No further details are available within the scope of Bellerophon, however, it does indicate that there has is a history of recombination within these sequences. The Maximum chi-squared test has been used to test the significance of the mosaics within the sequence. Putative recombinant sequence ST49 was compared putative parental sequences ST50 and ST15 and was found to have a mosaic of great significance ($p > 0.0001$). ST49 is a closely related lineage of ST50. The cutoff identified by the Maximum chi-squared test lies within the *clfB* sequences (between 3087 and 3203) where the mosaic begins and ST49 resembles the unrelated lineage ST15. The reconstruction of phylogeny for this species in chapter 3 enables the comparison of the relationships between lineages. This discerns that homology between ST49 and STs 1 and 15 have arisen as a result of recombination in *clfB* rather than descent.

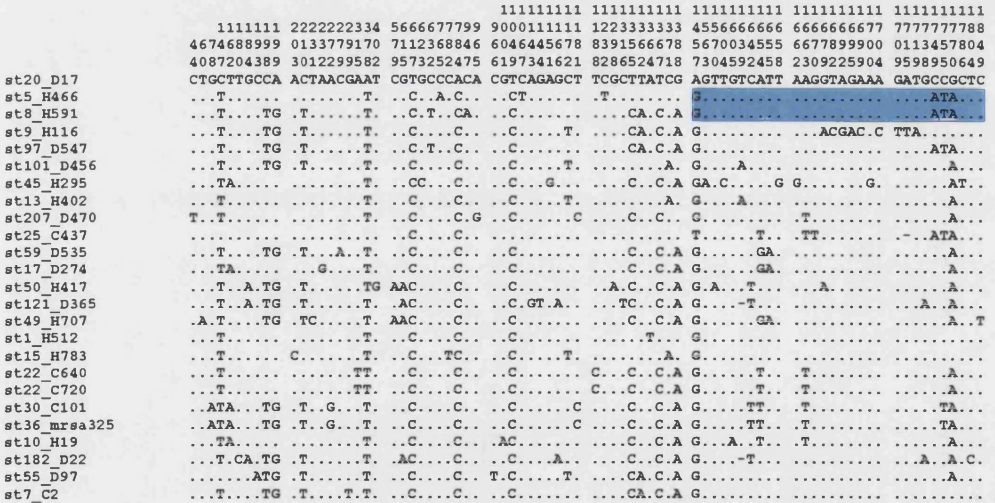


Figure 3. Polymorphic sites within the region of *arcC*, *crp* and *clfB*.....continued overleaf

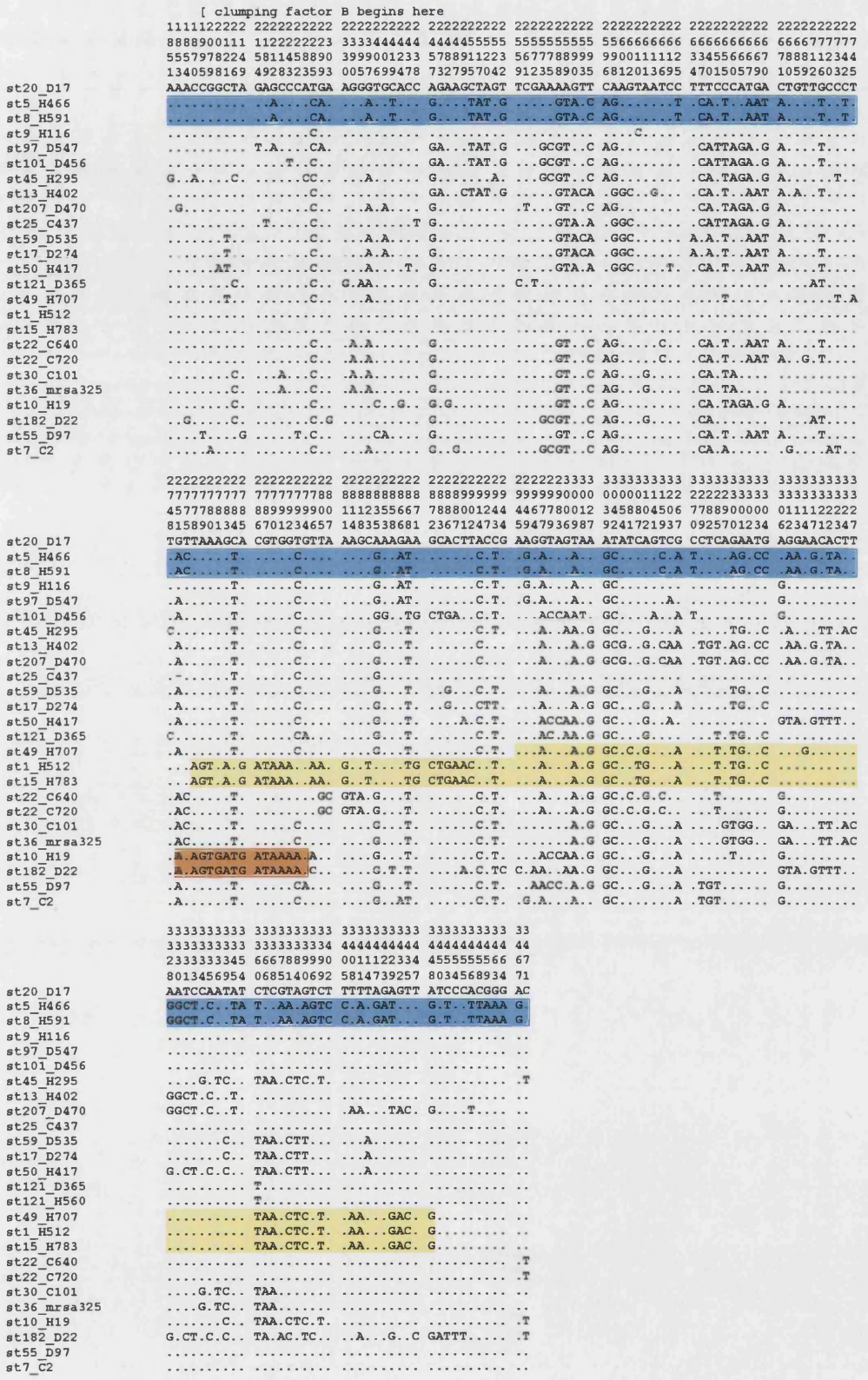


Figure 3 continued. Polymorphic sites within the region of *arcC*, *crp* and *clfB*.

6.2.2 Evidence for recombination within *clfB*

The accumulation of an excess of variation within the *clfB* gene has been demonstrated within Figures 2 and 3. We also find evidence for a higher incidence of recombination within the *clfB* locus compared to the *arcC* and *crp* genes. The minimum number of recombination events (R_M) has been calculated for the *arcC*, *crp* and *clfB* genes. This measure of recombination tells us that the R_M for the *arcC* and *crp* genes =1 whereas the R_M for *clfB* =34. Likewise, the Sawyer's Runs test finds no significant fragments in its analysis for the presence of recombinants for either *arcC* or *crp*. However, again significant recombination is detected within the *clfB* sequences (SSCF p= 0.0000 and SSUF p=0.0002).

We can also observe the impact of recombination in this region by comparing topologies of the individual genes within this region. In Figure 4 topological comparisons of neighbour-joining trees for *arcC*, *crp* and *clfB* are shown. Three examples of how relationships between strains representing different lineages can change between neighbouring genes are demonstrated. ST49 is a lineage of group 3 of the population, as defined in chapter 3. In the *arcC* and *crp* gene trees, members of this group, including ST49 are closely associated within the tree. However, within the *clfB* tree, this group cannot be resolved and lineages of this group are dispersed within the tree. Of note is ST49 sequence which is more closely associated with lineages ST1 and ST15 as a result of the significant mosaic identified by the Maximum chi-squared test and illustrated in figure 2. ST5 and ST8 are unrelated lineages which exist within group 2 of the *S. aureus* population. They are unassociated within the topologies of *arcC* and *crp*, yet in the *clfB* topology they are found on the same branch with complete sequence identity represented. In the same way there is homology between ST9 and ST20 represented in the *clfB* tree which is not observed within the *arcC* and *crp* trees.

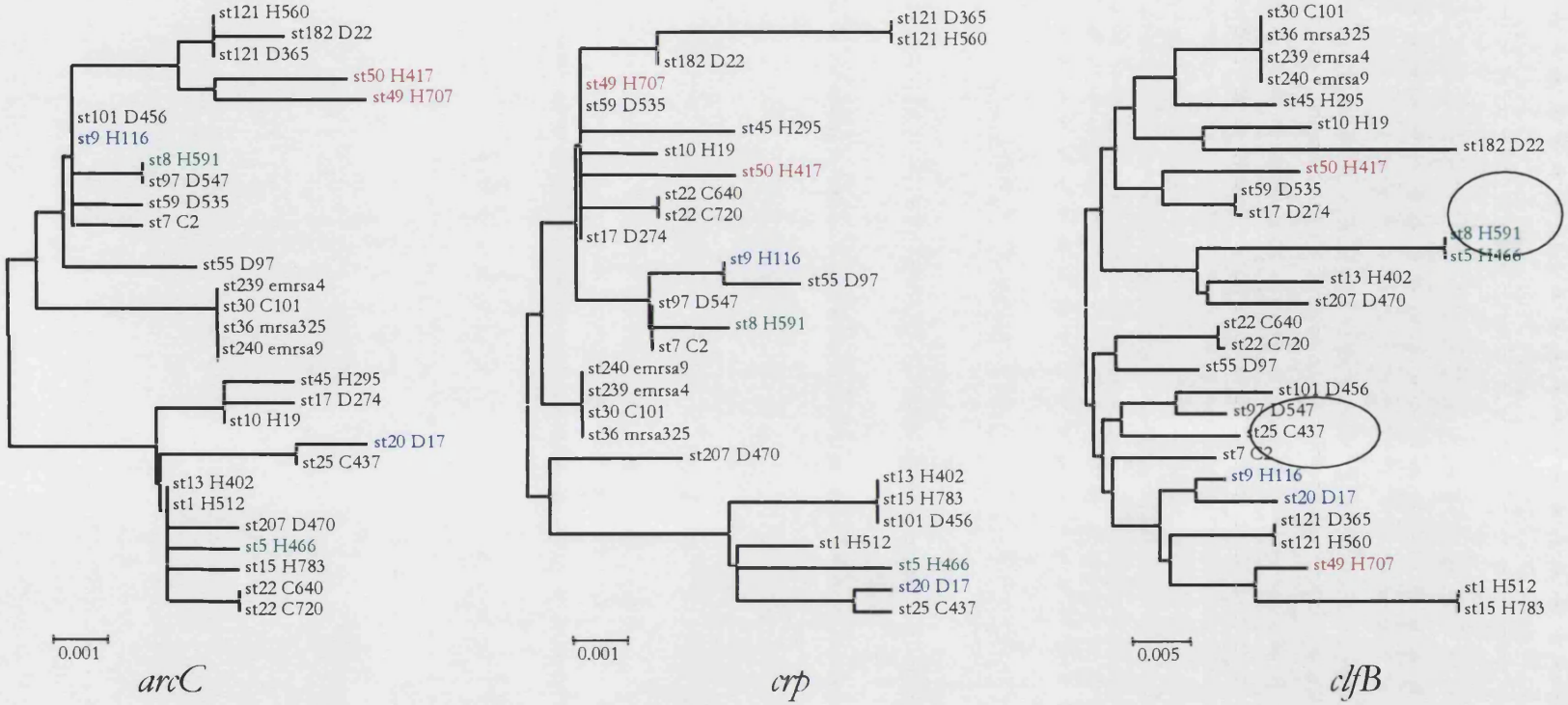
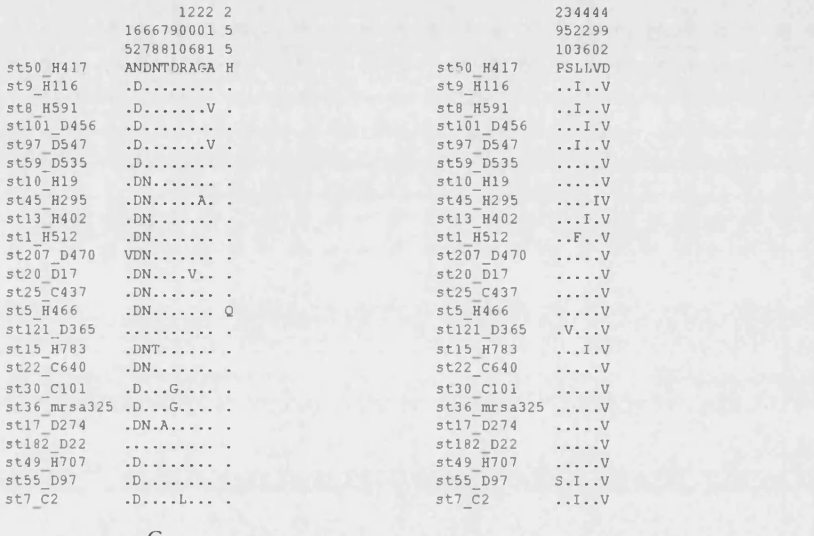


Figure 4. Neighbour joining trees for *arcC, crp* and *clfB* nucleotide sequences

The topology of the trees is different for each of the neighbouring genes reflecting the differences in variation found within these loci. Strains which appear closely related in one tree appear unrelated in another (indicated by coloured names). This incongruence provides further evidence that recombination has occurred within one or all of these loci.

Translation of the nucleotide sequences for this region reveals that variation seen within the sequences of *clfB* is not all synonymous (Figure 5). This may be interpreted as a reflection of selective pressures or relaxed functional constraint and will be considered further



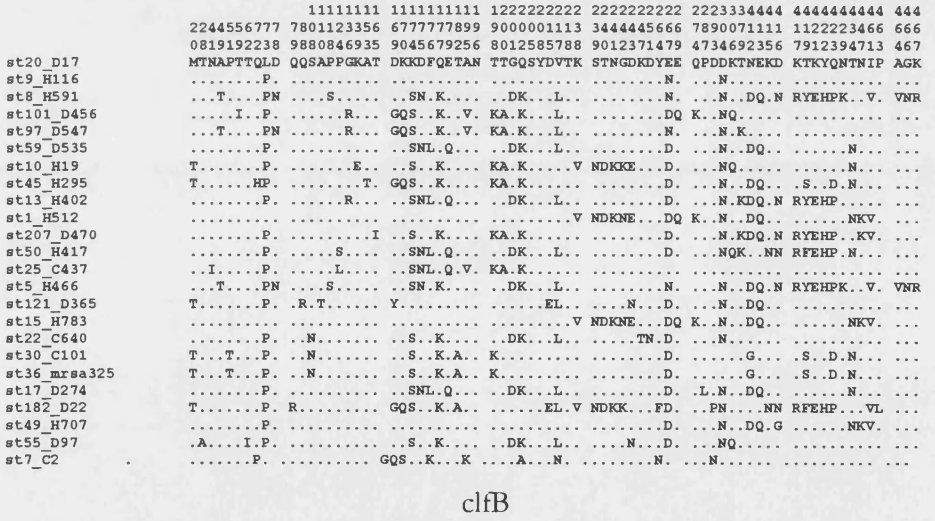


Figure 5. Partial protein sequences for arcC and clfB. The complete protein sequence for crp is shown.

6.2.3 Evidence for the role of selection

The d_S/d_N value for the sequences representing the 3 loci in this region have been calculated and are shown in table 1. d_S/d_N can be used as a measure of both functional constraint (the strength of purifying selection) and of positive selection. There is no evidence of positive selection ($d_S/d_N < 1$). The *crp* gene shows the greatest functional constraint whereas *clfB* has the most relaxed functional constraint. As a housekeeping gene *arcC* has a low value compared to most other housekeeping genes (Chapter 3, Table 1, page 75).

Genes	d_S	d_N	d_S/d_N
arcC	0.014	0.003	4.667
crp	0.016	0.002	8.000
clfB	0.067	0.021	3.190

Table 2. d_S/d_N for 3 genes in the sequence.

Note: the value for *arcC* is lower here than recorded in chapter 3 on account of the characterisation of a larger fragment within this study.

Although there is no evidence of positive selection acting upon any of the genes as a whole in this region, functional constraint and selective pressure may also vary between specific sites. The program Datamonkey (HYPHY package) is used to determine the selective pressures acting upon individual residues. No negatively or positively selected residues were detected within *arcC* and only one negatively selected residue was detected within *crp*. The A region of the clfB protein has been reported to exist in three subdomains which vary in their functional roles and capabilities (Perkins *et al.*, 2001). Datamonkey identifies 1 positively selected residue within the N2 subdomain and 45 further negatively selected residues within the *clfB* sequences (Figure 6).

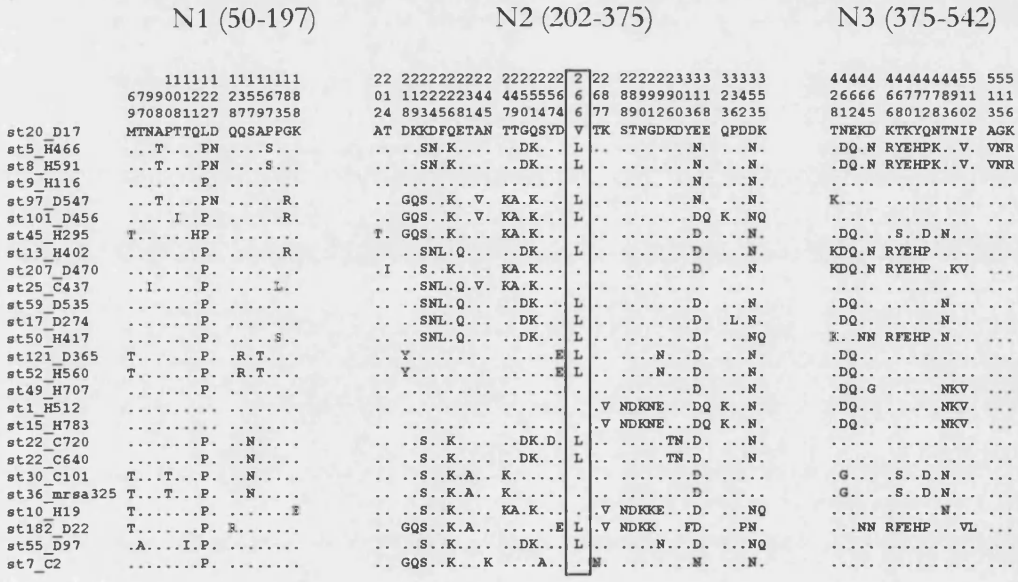


Figure 6. ClfB A region subdomains. The positively selected residue is highlighted in the box.

Differences in the topologies and functional constraint between individual subdomains of the clfB A region are shown in Figure 7. Even within a single domain of this protein, represented by three subdomains we find differences in the topologies and relationships between some strains. In subdomains N1 and N2 ST97 is closely associated with STs 5 and 8 yet is completely unrelated as represented by the topology for N3. Unrelated lineages ST49 and ST1 and 15 cluster in the topologies of subdomains N1 and N3 and yet in the central subdomain N2 ST49 is completely unassociated with STs 1 and 15. ST13 and ST207 are unrelated lineages of different population groups and this is reflected in the topology of N1. However, these two STs are very closely associated within the topologies of both N2 and N3 subdomains.

ArcC is the least congruent of the MLST genes compared to all other MLST genes. Interestingly, when we compare the three topologies of *arcC*, *crp* and *clfB* against the consensus tree generated in Chapter 3 using the SH test, *clfB* provides the highest likelihood of fit the consensus tree of 37 genes (Table 3).

Tree	SH score	Rank
consensus	best	
clfB	2182.439	1
crp	2472.511	2
arcC	2477.462	3

Table 3. SH scores for topologies of *arcC*, *crp* and *clfB*.

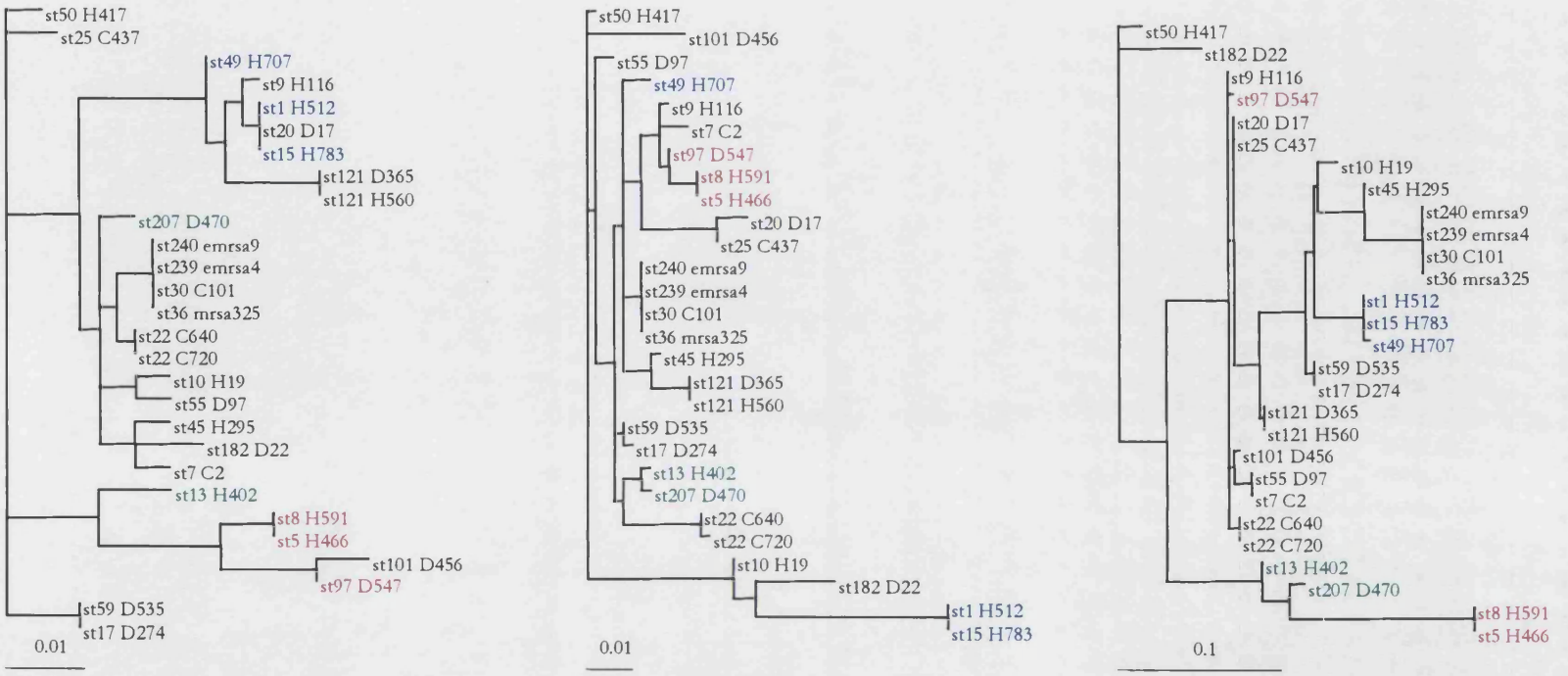


Figure 7. Changes in Maximum-likelihood tree topology over the A region subdomains of *clfB* alleles

Changes in relationships are illustrated by groups of strains (indicated by coloured groups of strain names). This incongruence provides further evidence that recombination has occurred within individual subdomains.

6.3 DISCUSSION

We find evidence for recombination within the A region of the SDR family member *clfB*. However, we find no evidence to suggest that the phylogenetic inconsistency observed for the *arcC* gene is a result of either recombination within *arcC* or hitchhiking of flanking recombinational events from the *clfB* gene. The hypothesis that incongruence in *arcC* is a result of proximity to *clfB* cannot be accepted. Recombination within the central gene of this region, *crp*, may have obscurred a *clfB* recombinant which flanked *arcC*. However, we find no evidence for extensive or repeated recombination within the *crp* gene to support this. In cases where a recombination event spans multiple loci there will be a trade-off between the relative selective pressures acting upon the genes involved. Any advantage conferred to the recombinant must be greater than the potential disadvantage conferred by variation within neighbouring loci. The hitchhiking of flanking DNA sequence into the *ddl* Housekeeping gene of *Streptococcus pneumoniae* has been reported. This has been attributed to the interspecific recombination of the neighbouring penicillin binding protein (pbp2b) conferring pencillin resistance to recombinants (Enright & Spratt, 1999). In this case, in the presence of penicillin, the selective advantage is a strong one and can be considered 'all or nothing' ie. resistance or sensitivity. Where the selective pressure is less clear cut as in examples of diversifying selection this may not be as strong a force for recombinants to be favoured if at the expense of neighbouring loci.

The results presented here are suggestive of small recombination events within the *clfB* gene. It may be more likely that small recombination events, which fall within single loci, will be less subject to purifying selection. Figure 7 shows the topologies within the subdomains of the A region only of *clfB* and how relationships between strains are changing even within a 1.4 Kb sequence. As found in Chapter 3, housekeeping genes have an intermediate level of purifying selection acting upon them compared to more relaxed purifying selection for ORPHANS and examples of high functional constraint for several information pathway genes. The variation within *clfB* is also inconsistent with the purging of variation at these loci as a result of a selective sweep. The gene residing between *arcC* and *clfB* may act as a buffer to the incorporation of *clfB* DNA sequence with flanking sequence. This gene in *Staphylococcus aureus* is similar to the *crp/fnr* family of

transcriptional regulators. With a role in the regulation of other loci, we find the highest d_S/d_N for the three genes for this gene, *crp*. However, this value could be considered more typical of a housekeeping gene rather than an information pathway gene when compared to the values obtained for these categories in Chapter 3 (Table 1, page 75).

It may be the case that there is a selective trade-off between these two neighbouring genes. In this way, *crp* may act as a 'buffer' since there is no evidence for recombination within this gene. A level of functional constraint of a neighbouring regulator may have inhibited the evolution of the housekeeping gene *arcC* although this seems unlikely to be a major factor since the measure of functional constraint at this locus is low.

An alternative explanation for incongruence at *arcC* is the presence of further potentially hypervariable genes upstream of *arcC*. The gene encoding the metalloprotease aureolysin is located approximately 5.5 Kb upstream of *arcC* with is neighbour *isaB* (immunodominant superantigen B) 7.8 Kb from *arcC*. The role of such loci in bacterial pathogenesis may be reflected in either hypervariability as a result of relaxed functional constraint or as a result of diversifying or positive selection pressures. As has been observed in other genes of 'accessory' function there may be have been recombination within these genes which could flank into neighbouring genes. However, as has been previously discussed, the patterns of variation across the region of *arcC*, *crp* and *clfB* and those seen for neighbouring genes in Chapter 3 (page 103), suggest that smaller events of recombination within the single genes are more frequent and more likely to persist within the population. It appears that the incongruence for *arcC* is in fact a result of a paucity of informative sites and in the absence of a clear explanation it appears that stochastic factors may be the determining factor. There is no evidence for recombination at this locus by an array of methods.

Even the highest scoring genes (by the SH test, Chapter 3) have a history of recombination which may typically introduce more diversity than point mutation alone, thus increasing the branch lengths and providing resolution comparable to that for the described phylogeny for this species (Chapter 3). This idea is supported by the fact that, despite recombination, *clfB* scores higher than *arcC* by the SH test when compared to the

consensus phylogeny for this species. The hypervariability of the *clfB* locus provides resolution of strains which is more informative, despite some inconsistencies, than that of *arcC*.

The characterisation and understanding of the mode of evolution at loci such as *clfB* is important due to their role in staphylococcal pathogenesis. There is no evidence in this study for extensive positive selection at this locus. This is consistent with the findings of only one positively selected site within sdr family member sdrE in a different putative subdomain of the A region (Chapter 5). It is unlikely that a single site would confer any selective advantage and I suggest that this finding is most likely to be reflective of an extreme relaxed functional constraint at this site. The finding of 45 negatively selected (conserved) sites also highlights the dynamic nature of a protein such as this. Even within a single domain there are extreme differences in the levels of functional constraint. Within the clfB protein there are variable numbers of serine-aspartate (SD) repeats and yet never fewer than are required to extend the protein through the cell wall and expose the A region. There is a conserved signal sequence and conserved LPXTG motif for peptidoglycan anchoring. The SLAVA motif is also conserved with proteolytic cleavage occuring between Ser^{197} and Leu^{198} and also between Ala^{199} and Val^{200} (McAleese *et al.*, 2001). The negatively selected sites presumably reflect the conservation of residues required for ligand binding and/or structural integrity.

Clumping factor B has been proposed as a potential vaccine candidate. Vaccination may play a significant role in the control strategies for *Staphylococcus aureus* in light of the rapid development of a broad range of antibiotic-resistance. A number of vaccines are being developed to provide both active and passive immunity. One of these includes a human immunoglobulin that is enriched for antibodies that recognize clumping factor A (Vernachio *et al.*, 2003). ClfB, sasG and WTA (wall teichoic acid) have been implicated as important in nasal colonisation (McAleese *et al.*, 2001; Roche *et al.*, 2003; Weidenmaier *et al.*, 2004) which is recognised as a significant risk factor for staphylococcal disease (von Eiff *et al.*, 2001). An increased understanding of the biology of *S. aureus* nasal colonisation could allow improved methods for controlling nasal and skin carriage. This understanding includes the mode of evolution of such loci. In this study the evolution of *clfB* by repeated

recombination has been demonstrated. The role of clfB in nasal colonisation makes implicates this protein as a potential vaccine candidate. However, vaccine design must recognise the mode of evolution at this locus. There is no evidence for strong or positive diversifying selection at this locus, however with a change in environmental conditions such as the introduction of a vaccine the propensity for recombination at this locus may be problematic.

OVERVIEW AND CONCLUDING REMARKS

In this thesis I have examined the relevance of gene function to recombination rates and phylogenetic reliability. This study shows that there is no association between functional constraint (d_S/d_N) and gene category. Perhaps of greater significance is the finding that both homologous recombination and phylogenetic reliability do not appear to show a simple relationship to functional constraint. In particular, evidence has been presented for the compatibility of recombination within genes which are likely to be under a high functional constraint. Synonymous replacements will be subject to little, if any, purifying selection since natural selection primarily acts at the protein level. These findings challenge the 'complexity hypothesis' which proposes that epistatic interactions limit recombination within information pathway genes. Therefore, although these genes represent the essential 'core' set of genes, at least at the intraspecies level, they are not necessarily the most reliable phylogenetic markers. The low degree of sequence diversity within many of these genes also limits their use for phylogenetic analysis.

It is relevant to speculate on the likely selective outcomes of recombination events and point mutations. On the one hand, a single recombinational replacement involves the substitution of many nucleotide sites within a coding sequence, whereas a point mutation will only change a single site. Intuitively, we might therefore expect recombination events to impose a greater selective cost than point mutations. However, to counter this, we must also consider the fact that recombination events will involve sequences which already exist within the population, and will have therefore already passed a selective filter, whereas *de novo* point mutations will result in novel changes which may have a more dramatic effect on protein function. It is noteworthy that many of the clear mosaics noted within the data generated in this study consist largely of synonymous changes, at least in the core genes (e.g. there is only one nonsynonymous substitution in significant mosaics of 25 and 43 variable sites, in the gene *hutI*), and this points to the possibility that recombination events are more likely to be selectively tolerated, on the whole, than point mutations. This observation also potentially provides a further criterion for the identification of recombination events, as runs of *largely synonymous* polymorphic sites.

By examining the relationship between phylogenetic reliability and sequence diversity (Chapter 3, Figure 12), it is clear that there is an optimal window of variation. Too few informative sites within a coding sequence results in the clustering of strains which are clearly distinguishable on the basis of other gene loci, whereas in contrast too much variation results in the separation of closely related strains. The use of housekeeping genes for MLST in *S. aureus* is appropriate, as this species is highly clonal, and almost any gene locus will assign a given strain to the same clonal lineage. Indeed, the utility of the current MLST scheme has been proven by it use in the reconstruction of short-term evolutionary relationships within clonal complexes. However, the data also indicate that the MLST genes are perhaps not ideal for resolving longer-term evolutionary relationships, and other sets of genes appear to perform better in this regard. Phylogenetic inconsistency within these genes can be accounted for by a paucity of informative sites rather than high recombination or proximity to hypervariable loci (Chapter 6), and larger fragments of these genes, or a greater sample of gene loci, would be more appropriate for detailed phylogenetic study.

Although recombination has been detected in all gene categories a common phylogenetic signal is still apparent when large amounts of sequence data, from 37 separate loci, are combined. This dataset provides a greatly improved phylogeny than that based on the original 7 MLST housekeeping genes (Feil *et al.*, 20013. The major divisions within the population revealed by these data are generally consistent with those reported by Melles *et al.*, 2004 with the caveat that the samples used in the two studies differ in the emphasis placed on each of the major groups. However, the evolutionary and ecological significance of these divisions within the population are unclear, although it is possible that the major division between Group 1 and Group 2 strains, which is even apparent from the MLST data (Feil *et al.*, 2003), represents an ancient split in the population. This division potentially provides some justification for the subdivision of *S. aureus* into two sub-species, although such a distinction would require further evidence from phenotypic, as well as genotypic, studies.

No particular gene category was found to be more phylogenetically reliable than any other, although both consensus informational pathway and MLST housekeeping phylogenies

scored poorly against the consensus tree. Surprisingly, the best scoring gene was identified as *sasF*, a surface exposed protein-coding gene, which intuitively might be expected to have performed rather poorly. No clear parameters were identified that could systematically predict the performance of single loci in terms of phylogenetic reliability. This implicates stochastic forces as possibly a more important determinant of phylogenetic reliability than selection.

The reconstruction of the intraspecific phylogeny for *S. aureus* has provided a detailed picture of the distribution of allelic variants of the accessory SDR family member *sdrE*. It is clear that there has been a history of both gene transfer and allelic replacement within this locus. The functional redundancy, and possible multiple roles, of Staphylococcal MSCRAMMs makes the implications of the loss or acquisition of such a locus unclear. Virulence potential in the Staphylococcus results from the complex interactions of many bacterial and host gene products, and it is therefore over-simplistic to imagine particular strains as either virulent or non-virulent. Despite these difficulties, Peacock *et al.*, 2002 managed to demonstrate an association between the number of virulence-associated determinants, including sdrE, and the propensity to cause disease.

There are two possible scenarios to account for a higher virulence potential in strains containing *sdrE*. Firstly, sdrE acts as an adhesin and facilitates cell binding to host tissues. It is possible that aggressive colonisation in strains containing this gene leads to more localised tissue damage, and therefore an increased likelihood of access to the blood stream and disseminated infection. Secondly, the role of sdrE in platelet aggregation may infer an increased risk of endocarditis through the formation of thrombi around the heart valves. These two possibilities are not mutually exclusive, and both the adhesion to host tissues and platelet aggregation may play a significant role.

It is likely that there are further roles for this protein which are as yet uncharacterised but the evidence presented here suggests that there is little functional difference between the two allelic variants found at the *sdrE* locus. However, we have seen that the replacement of diverse alleles producing diverse proteins, has not affected their hosts ability to activate the aggregation of platelets in human hosts. Furthermore, evidence for gene transfer

(rather than homologous recombination) of a locus positively associated with invasive disease has implications for the emergence and maintenance of disease potential in this species. The intraspecifc transfer of disease-associated determinants, and increasing drug-resistance in *Staphylococcus aureus,* paints a worrying picture and underscores the difficulties in vaccine development. Members of the MSCRAMM family of surface exposed proteins, such as the ubiquitous clumping factor B (clfB), may be suitable candidates but the variation within these loci and the propensity for recombination must be taken into account if vaccination is to be an effective alternative to antibiotic therapy.

LITERATURE CITED

Anon (1995). Epidemic methicillin-resistant *Staphylococcus aureus*. *CDR* **5**, 35.

Aris-Brosou, S. (2005). Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. *Mol Biol Evol* **22**, 200-209.

Barber, M. & Rozwadowska-Dowzenko, M. (1948). Infection by penicillin-resistant staphylococci. *Lancet* ii, 641-644.

Bandelt, H. J. & Dress, A. W. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* **1**, 242-252.

Boelaert, J. R., Van Landuyt, H. W., Godard, C. A. & other authors (1993). Nasal mupirocin ointment decreases the incidence of *Staphylococcus aureus* bacteraemias in haemodialysis patients. *Nephrol Dial Transplant* **8**, 235-239.

Bergdoll, M. S., Crass, B. A., Reiser, R. F., Robbins, R. N. & Davis, J. P. (1981). A new staphylococcal enterotoxin, enterotoxin F, associated with toxic-shock-syndrome *Staphylococcus aureus* isolates. *Lancet* **1**, 1017-1021.

Brown, D. F. & Reynolds, P. E. (1980). Intrinsic resistance to beta-lactam antibiotics in *Staphylococcus aureus*. *FEBS Lett* **122**, 275-278.

Bruck, I. & O'Donnell, M. (2000). The DNA replication machine of a gram-positive organism. *J Biol Chem* **275**, 28971-28983.

CDC (2002a). Centers for Disease control and Prevention. Public Health Dispatch: vancomycin-resistant Staphylococcus aureus- Pennsylvania. *Morb Mortal Wkly* **51**, 902-903.

CDC (2002b). Centers for Disease Control and Prevention. Staphylococcus aureus resistant to vancomycin- United States. *Morb Mortal Wkly* **51**, 565-567.

LITERATURE CITED

CDC (2004). Centers for Disease control and prevention. Vancomycin-resistant Staphylococcus aureus-New York. *Morb Mortal Wkly* **53**, 322-323.

Chilton, M. D., Drummond, M. H., Merio, D. J., Sciaky, D., Montoya, A. L., Gordon, M. P. & Nester, E. W. (1977). Stable incorporation of plasmid DNA into higher plant cells: the molecular basis of crown gall tumorigenesis. *Cell* **11**, 263-271.

Coggins, L. W. & O'Prey, M. (1989). DNA tertiary structures formed in vitro by misaligned hybridization of multiple tandem repeat sequences. *Nucleic Acids Res* **17**, 7417-7426.

Cohan, F. M. (1995). Does recombination constrain neutral divergence among bacterial taxa? *Evolution.* **49** (1), 164-175.

Cui, L., Murakami, H., Kuwahara-Arai, K., Hanaki, H. & Hiramatsu, K. (2000). Contribution of a thickened cell wall and its glutamine nonamidated component to the vancomycin resistance expressed by *Staphylococcus aureus* Mu50. *Antimicrob Agents Chemother* **44**, 2276-2285.

Daubin, V., Gouy, M. & Perriere, G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* **12**, 1080-1090.

Deppenmeier, U., Johann, A., Hartsch, T. & other authors (2002). The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol* **4**, 453-461.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129.

Downer, R., Roche, F., Park, P. W., Mecham, R. P. & Foster, T. J. (2002). The elastin-binding protein of *Staphylococcus aureus* (EbpS) is expressed at the cell surface as an integral membrane protein and not as a cell wall-associated protein. *J Biol Chem* **277**, 243-250.

Drouin, G., Prat, F., Ell, M. & Clarke, G. D. (1999). Detecting and characterizing gene conversions between multigene family members. *Mol Biol Evol* **16**, 1369-1390.

Dufour, P., Jarraud, S., Vandenesch, F., Greenland, T., Novick, R. P., Bes, M., Etienne, J. & Lina, G. (2002). High genetic variability of the agr locus in Staphylococcus species. *J Bacteriol* **184**, 1180-1186.

Duthie, E. S. & Lorenz, L. L. (1952). Staphylococcal coagulase; mode of action and antigenicity. *J Gen Microbiol* **6**, 95-107.

Eisen, J. A. (2000). Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev* **10**, 606-611.

Enright, M. C. & Spratt, B. G. (1999). Extensive variation in the ddl gene of penicillin-resistant *Streptococcus pneumoniae* results from a hitchhiking effect driven by the penicillin-binding protein 2b gene. *Mol Biol Evol* **16**, 1687-1695.

Enright, M. C., Day, N. P., Davies, C. E., Peacock, S. J. & Spratt, B. G. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *J Clin Microbiol* **38**, 1008-1015.

Enright, M. C., Robinson, D. A., Randle, G., Feil, E. J., Grundmann, H. & Spratt, B. G. (2002). The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc Natl Acad Sci U S A* **99**, 7687-7692.

Espersen, F. & Clemmensen, I. (1982). Isolation of a fibronectin-binding protein from *Staphylococcus aureus*. *Infect Immun* **37**, 526-531.

Espersen, F., Clemmensen, I. & Barkholt, V. (1985). Isolation of *Staphylococcus aureus* clumping factor. *Infect Immun* **49**, 700-708.

LITERATURE CITED

Falush, D., Kraft, C., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M. & Suerbaum, S. (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci U S A* **98**, 15056-15061.

Fearnhead, P. & Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299-1318.

Feil, E. J., Holmes, E. C., Bessen, D. E. & other authors (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci U S A* **98**, 182-187.

Feil, E. J., Cooper, J. E., Grundmann, H. & other authors (2003). How clonal is *Staphylococcus aureus*? *J Bacteriol* **185**, 3307-3316.

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P. & Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**, 1518-1530.

Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R. & Musser, J. M. (2001). Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc Natl Acad Sci U S A* **98**, 8821-8826.

Garcia-Vallve, S., Romeu, A. & Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10**, 1719-1725.

Gasson, M. J. (1983). Plasmid complements of *Streptococcus lactis* NCDO 712 and other lactic streptococci after protoplast-induced curing. *J Bacteriol* **154**, 1-9.

Gillespie, J. H. (1991). The causes of molecular evolution. Oxford Univ. Press, New York.

LITERATURE CITED

Goodman, S. D. & Scocca, J. J. (1988). Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc Natl Acad Sci U S A* **85**, 6982-6986.

Greene, C., McDevitt, D., Francois, P., Vaudaux, P. E., Lew, D. P. & Foster, T. J. (1995). Adhesion properties of mutants of *Staphylococcus aureus* defective in fibronectin-binding proteins and studies on the expression of fnb genes. *Mol Microbiol* **17**, 1143-1152.

Grundmann, H., Tami, A., Hori, S., Halwani, M. & Slack, R. (2002). Nottingham Staphylococcus aureus population study: prevalence of MRSA among elderly people in the community. *British Medical Journal* **324**, 1365-1366.

Haldane, J, B, S. (1957). The cost of natural selection. *J Gen.* **55**. 511-524

Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. (2003). The genetic core of the universal ancestor. *Genome Res* **13**, 407-412.

Harris-Warrick R. M., L. J. (1978). Interspecies transformation in Bacillus: mechansim of heterologous intergenote transformation. *J Bacteriol* **133**, 1246-1253.

Hartford, O., Francois, P., Vaudaux, P. & Foster, T. J. (1997). The dipeptide repeat region of the fibrinogen-binding protein (clumping factor) is required for functional expression of the fibrinogen-binding domain on the *Staphylococcus aureus* cell surface. *Mol Microbiol* **25**, 1065-1076.

Hartford, O., O'Brien, L., Schofield, K., Wells, J. & Foster, T. J. (2001a). The Fbe (SdrG) protein of *Staphylococcus epidermidis* HB promotes bacterial adherence to fibrinogen. *Microbiology* **147**, 2545-2552.

Hartford, O. M., Wann, E. R., Hook, M. & Foster, T. J. (2001b). Identification of residues in the *Staphylococcus aureus* fibrinogen-binding MSCRAMM clumping factor A (ClfA) that are important for ligand binding. *J Biol Chem* **276**, 2466-2473.

LITERATURE CITED

Hartleib, J., Kohler, N., Dickinson, R. B. & other authors (2000). Protein A is the von Willebrand factor binding protein on *Staphylococcus aureus*. *Blood* **96**, 2149-2156.

Hartman, B. J. & Tomasz, A. (1984). Low-affinity penicillin-binding protein associated with beta-lactam resistance in *Staphylococcus aureus*. *J Bacteriol* **158**, 513-516.

Hiramatsu, K., Aritaka, N., Hanaki, H., Kawasaki, S., Hosoda, Y., Hori, S., Fukuchi, Y. & Kobayashi, I. (1997). Dissemination in Japanese hospitals of strains of *Staphylococcus aureus* heterogeneously resistant to vancomycin. *Lancet* **350**, 1670-1673.

Hiramatsu, K., Okuma, K., Ma, X. X., Yamamoto, M., Hori, S. & Kapi, M. (2002). New trends in *Staphylococcus aureus* infections: glycopeptide resistance in hospital and methicillin resistance in the community. *Curr Opin Infect Dis* **15**, 407-413.

Hoffmaster, A. R., Fitzgerald, C. C., Ribot, E., Mayer, L. W. & Popovic, T. (2002). Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak, United States. *Emerg Infect Dis* **8**, 1111-1116.

Holden, M. T., Feil, E. J., Lindsay, J. A. & other authors (2004). Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S A* **101**, 9786-9791.

Huber, T., Faulkner, G. & Hugenholtz, P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317-2319.

Hudson, R. R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147-164.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics* **159**, 1805-1817.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755.

Hugenholtzt, P. & Huber, T. (2003). Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int J Syst Evol Microbiol* **53**, 289-293.

Hughes, A. L. & Friedman, R. (2005). Nucleotide substitution and recombination at orthologous loci in *Staphylococcus aureus*. *J Bacteriol* **187**, 2698-2704.

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68-73.

Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* ***96***, 3801-3806.

Janzon, L. & Arvidson, S. (1990). The role of the delta-lysin gene (hld) in the regulation of virulence genes by the accessory gene regulator (agr) in *Staphylococcus aureus*. *Embo J* **9**, 1391-1399.

Jarraud, S., Lyon, G. J., Figueiredo, A. M., Gerard, L., Vandenesch, F., Etienne, J., Muir, T. W. & Novick, R. P. (2000). Exfoliatin-producing strains define a fourth agr specificity group in *Staphylococcus aureus*. *J Bacteriol* **182**, 6517-6522.

Ji, G., Beavis, R. & Novick, R. P. (1997). Bacterial interference caused by autoinducing peptide variants. *Science* **276**, 2027-2030.

Jevons, M. P. (1961). Celbenin-resistant staphylococci. *British Medical Journal* i, 124-125.

Jonsson, K., Signas, C., Muller, H. P. & Lindberg, M. (1991). Two different genes encode fibronectin binding proteins in *Staphylococcus aureus*. The complete nucleotide sequence and characterization of the second gene. *Eur J Biochem* **202**, 1041-1048.

LITERATURE CITED

Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. (2002). Microevolutionary genomics of bacteria. *Theor Popul Biol* ***61***, 435-447.

Josefsson, E., McCrea, K. W., Ni Eidhin, D., O'Connell, D., Cox, J., Hook, M. & Foster, T. J. (1998). Three new members of the serine-aspartate repeat protein multigene family of *Staphylococcus aureus*. *Microbiology* ***144*** (Pt 12), 3387-3395.

Karlin, S. & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* ***11***, 283-290.

Karlin, S., Mrazek, J. & Campbell, A. M. (1998). Codon usages in different gene classes of the *Escherichia coli* genome. *Mol Microbiol* ***29***, 1341-1355.

Karlin, S. & Mrazek, J. (2000). Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol* ***182***, 5238-5250.

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet Res* ***11***, 247-269.

King, J. L. & Jukes, T. H. (1969). Non-Darwinian evolution. *Science* ***164***, 788-798.

Klein, J. L., Petrovic, Z., Treacher, D. & Edgeworth, J. (2003). Severe community-acquired pneumonia caused by Panton-Valentine leukocidin-positive *Staphylococcus aureus*: first reported case in the United Kingdom. *Intensive Care Med* **29**, 1399.

Kluytmans, J. (1998). Reduction of surgical site infections in major surgery by elimination of nasal carriage of *Staphylococcus aureus*. *J Hosp Infect* **40** Suppl B, S25-29.

Kosakovsky Pond, S. L., Frost, S. D. & Muse, S. V. (2004). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*.

LITERATURE CITED

Kraulis, P. J. (1991). "MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures", Journal of Applied Crystallography **24**, 946-950.

Kreiswirth, B., Kornblum, J., Arbeit, R. D., Eisner, W., Maslow, J. N., McGeer, A., Low, D. E. & Novick, R. P. (1993). Evidence for a clonal origin of methicillin resistance in *Staphylococcus aureus*. *Science* **259**, 227-230.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244-1245.

Kunst, F., Ogasawara, N., Moszer, I. & other authors (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249-256.

Kurland, C. G. (2000). Something for everyone. Horizontal gene transfer in evolution. *EMBO Rep* **1**, 92-95.

Kuroda, M., Ohta, T., Uchiyama, I. & other authors (2001). Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* **357**, 1225-1240.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680-685.

Lawrence, J. G. & Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**, 383-397.

Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A* **95**, 9413-9417.

Lawrence, J. G. & Ochman, H. (2002). Reconciling the many faces of lateral gene transfer. *Trends Microbiol* **10**, 1-4.

Levin, B. R. (1981). Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* **99**, 1-23.

Levin, B. R. (1988). Frequency-dependent selection in bacterial populations. *Philos Trans R Soc Lond B Biol Sci* **319**, 459-472.

Lina, G., Piemont, Y., Godail-Gamot, F., Bes, M., Peter, M. O., Gauduchon, V., Vandenesch, F. & Etienne, J. (1999). Involvement of Panton-Valentine leukocidin-producing *Staphylococcus aureus* in primary skin infections and pneumonia. *Clin Infect Dis* **29**, 1128-1132.

Lindsay, J. A., Holden, M. T. G. (2004). *Staphylococcus aureus*: superbug, super genome? *Trends Microbiol* **12**, 378-385.

Logsdon, J. M. & Faguy, D. M. (1999). Thermotoga heats up lateral gene transfer. *Curr Biol* **9**, R747-751.

Luzar, M. A., Coles, G. A., Faller, B. & other authors (1990). *Staphylococcus aureus* nasal carriage and infection in patients on continuous ambulatory peritoneal dialysis. *N Engl J Med* **322**, 505-509.

Maiden, M. C., Bygraves, J. A., Feil, E. & other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140-3145.

Majewski, J. & Cohan, F. M. (1998). The effect of mismatch repair and heteroduplex formation on sexual isolation in Bacillus. *Genetics* **148**, 13-18.

Majewski, J. & Cohan, F. M. (1999). DNA sequence similarity requirements for interspecific recombination in Bacillus. *Genetics* **153**, 1525-1533.

Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatusov, R. L., Wolf, Y. I. & Koonin, E. V. (1999). Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res* **9**, 608-628.

McAleese, F. M., Walsh, E. J., Sieprawska, M., Potempa, J. & Foster, T. J. (2001). Loss of clumping factor B fibrinogen binding activity by *Staphylococcus aureus* involves cessation of transcription, shedding and cleavage by metalloprotease. *J Biol Chem* **276**, 29969-29978.

McDevitt, D., Francois, P., Vaudaux, P. & Foster, T. J. (1994). Molecular characterization of the clumping factor (fibrinogen receptor) of *Staphylococcus aureus*. *Mol Microbiol* **11**, 237-248.

McDevitt, D. & Foster, T. J. (1995). Variation in the size of the repeat region of the fibrinogen receptor (clumping factor) of *Staphylococcus aureus strains*. *Microbiology* **141** (Pt 4), 937-943.

McDevitt, D., Francois, P., Vaudaux, P. & Foster, T. J. (1995). Identification of the ligand-binding domain of the surface-located fibrinogen receptor (clumping factor) of *Staphylococcus aureus*. *Mol Microbiol* **16**, 895-907.

McDevitt, D., Nanavaty, T., House-Pompeo, K., Bell, E., Turner, N., McIntire, L., Foster, T. & Hook, M. (1997). Characterization of the interaction between the *Staphylococcus aureus* clumping factor (ClfA) and fibrinogen. *Eur J Biochem* **247**, 416-424.

Melles, D. C., Gorkink, R. F., Boelens, H. A. & other authors (2004). Natural population dynamics and expansion of pathogenic clones of *Staphylococcus aureus*. *J Clin Invest* **114**, 1732-1740.

Milkman, R. & Bridges, M. M. (1990). Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* **126**, 505-517.

Milkman, R., Jaeger, E. & McBride, R. D. (2003). Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. *Genetics* **163**, 475-483.

Misawa, K. & Nei, M. (2003). Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. *J Mol Evol* **57** Suppl 1, S290-296.

Morfeldt, E., Janzon, L., Arvidson, S. & Lofdahl, S. (1988). Cloning of a chromosomal locus (exp) which regulates the expression of several exoprotein genes in *Staphylococcus aureus*. *Mol Gen Genet* **211**, 435-440.

Moszer, I., Rocha, E. P. & Danchin, A. (1999). Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr Opin Microbiol* **2**, 524-528.

Musser, J. M., Schlievert, P. M., Chow, A. W., Ewan, P., Kreiswirth, B. N., Rosdahl, V. T., Naidu, A. S., Witte, W. & Selander, R. K. (1990). A single clone of *Staphylococcus aureus* causes the majority of cases of toxic shock syndrome. *Proc Natl Acad Sci U S A* **87**, 225-229.

Musser, J. M. & Kapur, V. (1992). Clonal analysis of methicillin-resistant *Staphylococcus aureus* strains from intercontinental sources: association of the mec gene with divergent phylogenetic lineages implies dissemination by horizontal transfer and recombination. *J Clin Microbiol* **30**, 2058-2063.

Nesbo, C. L., Boucher, Y. & Doolittle, W. F. (2001). Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. *J Mol Evol* **53**, 340-350.

Nguyen, M. H., Kauffman, C. A., Goodman, R. P., Squier, C., Arbeit, R. D., Singh, N., Wagener, M. M. & Yu, V. L. (1999). Nasal carriage of and infection with *Staphylococcus aureus* in HIV-infected patients. *Ann Intern Med* **130**, 221-225.

Ni Eidhin, D., Perkins, S., Francois, P., Vaudaux, P., Hook, M. & Foster, T. J. (1998). Clumping factor B (ClfB), a new surface-located fibrinogen-binding adhesin of *Staphylococcus aureus*. *Mol Microbiol* **30**, 245-257.

Noble, W. C., Virani, Z. & Cree, R. G. (1992). Co-transfer of vancomycin and other resistance genes from *Enterococcus faecalis* NCTC 12201 to *Staphylococcus aureus*. *FEMS Microbiol Lett* **72**, 195-198.

Novick, R. P., Ross, H. F., Projan, S. J., Kornblum, J., Kreiswirth, B. & Moghazeh, S. (1993). Synthesis of staphylococcal virulence factors is controlled by a regulatory RNA molecule. *Embo J* **12**, 3967-3975.

Novick, R. P. & Muir, T. W. (1999). Virulence gene regulation by peptides in staphylococci and other Gram-positive bacteria. *Curr Opin Microbiol* **2**, 40-45.

O'Brien, L., Kerrigan, S. W., Kaw, G., Hogan, M., Penades, J., Litt, D., Fitzgerald, D. J., Foster, T. J. & Cox, D. (2002a). Multiple mechanisms for the activation of human platelet aggregation by *Staphylococcus aureus*: roles for the clumping factors ClfA and ClfB, the serine-aspartate repeat protein SdrE and protein A. *Mol Microbiol* **44**, 1033-1044.

O'Brien, L. M., Walsh, E. J., Massey, R. C., Peacock, S. J. & Foster, T. J. (2002b). *Staphylococcus aureus* clumping factor B (ClfB) promotes adherence to human type I cytokeratin 10: implications for nasal colonization. *Cell Microbiol* **4**, 759-770.

O'Connell, D. P., Nanavaty, T., McDevitt, D., Gurusiddappa, S., Hook, M. & Foster, T. J. (1998). The fibrinogen-binding MSCRAMM (clumping factor) of *Staphylococcus aureus* has a Ca2+-dependent inhibitory site. *J Biol Chem* **273**, 6821-6829.

Ogston, A. (1882). *Micrococcus* poisoning. *J Anat* **17**, 24-58.

Palma, M., Haggar, A. & Flock, J. I. (1999). Adherence of *Staphylococcus aureus* is enhanced by an endogenous secreted protein with broad binding activity. *J Bacteriol* **181**, 2840-2845.

Parkhill, J., Sebaihia, M., Preston, A. & other authors (2003). Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* **35**, 32-40.

Patti, J. M., Jonsson, H., Guss, B., Switalski, L. M., Wiberg, K., Lindberg, M. & Hook, M. (1992). Molecular characterization and expression of a gene encoding a *Staphylococcus aureus* collagen adhesin. *J Biol Chem* **267**, 4766-4772.

Peacock, S. J., de Silva, I. & Lowy, F. D. (2001). What determines nasal carriage of *Staphylococcus aureus*? *Trends Microbiol* **9**, 605-610.

Peacock, S. J., Moore, C. E., Justice, A., Kantzanou, M., Story, L., Mackie, K., O'Neill, G. & Day, N. P. (2002). Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. *Infect Immun* **70**, 4987-4996.

Peng, H. L., Novick, R. P., Kreiswirth, B., Kornblum, J. & Schlievert, P. (1988). Cloning, characterization, and sequencing of an accessory gene regulator (agr) in *Staphylococcus aureus*. *J Bacteriol* **170**, 4365-4372.

Perkins, S., Walsh, E. J., Deivanayagam, C. C., Narayana, S. V., Foster, T. J. & Hook, M. (2001). Structural organization of the fibrinogen-binding region of the clumping factor B MSCRAMM of *Staphylococcus aureus*. *J Biol Chem* **276**, 44721-44728.

Ponnuraj, K., Bowden, M. G., Davis, S., Gurusiddappa, S., Moore, D., Choe, D., Xu, Y., Hook, M. & Narayana, S. V. (2003). A "dock, lock, and latch" structural model for a staphylococcal adhesin binding to fibrinogen. *Cell* **115**, 217-228.

Prevost, G., Mourey, L., Colin, D. A. & Menestrina, G. (2001). Staphylococcal pore-forming toxins. *Curr Top Microbiol Immunol* **257**, 53-83.

Rayssiguier, C., Thaler, D. S. & Radman, M. (1989). The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* **342**, 396-401.

Recsei, P., Kreiswirth, B., O'Reilly, M., Schlievert, P., Gruss, A. & Novick, R. P. (1986). Regulation of exoprotein gene expression in *Staphylococcus aureus* by agar. Mol Gen Genet **202**, 58-61.

Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K. & Whittam, T. S. (2000). Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**, 64-67.

Reeves, P. (1993). Evolution of Salmonella O antigen variation by interspecific gene transfer on a large scale. *Trends Genet* **9**, 17-22.

Reeves, P. (1995). Role of O-Antigen variation in the immune response. *Trends Microbiol* **3**, 381-386.

Robinson, D. A. & Enright, M. C. (2004). Evolution of *Staphylococcus aureus* by large chromosomal replacements. *J Bacteriol* **186**, 1060-1064.

Robinson, D. A., Monk, A. B., Cooper, J. E., Feil, E. J. and Enright, M. C. Evidence for Recombination and Incomplete Lineage sorting at the Accessory Gene Regulator (agr) locus in *Staphylococcus aureus*. submitted.

Rocha, E. P. C., Maynard Smith, J., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H and Feil, E. J. Changes in the d_S/d_N ratio during intraspecies divergence in bacteria. *J Theor Biol.* in press

Roche, F. M., Massey, R., Peacock, S. J., Day, N. P., Visai, L., Speziale, P., Lam, A., Pallen, M. & Foster, T. J. (2003a). Characterization of novel LPXTG-containing proteins of *Staphylococcus aureus* identified from genome sequences. *Microbiology* **149**, 643-654.

Roche, F. M., Meehan, M. & Foster, T. J. (2003b). The *Staphylococcus aureus* surface protein SasG and its homologues promote bacterial adherence to human desquamated nasal epithelial cells. *Microbiology* **149**, 2759-2767.

Roche, F. M., Downer, R., Keane, F., Speziale, P., Park, P. W. & Foster, T. J. (2004). The N-terminal A domain of fibronectin-binding proteins A and B promotes adhesion of *Staphylococcus aureus* to elastin. *J Biol Chem* **279**, 38433-38440.

Sabat, A., Krzyszton-Russjan, J., Strzalka, W., Filipek, R., Kosowska, K., Hryniewicz, W., Travis, J. & Potempa, J. (2003). New method for typing Staphylococcus aureus strains: multiple-locus variable-number tandem repeat analysis of polymorphism and genetic relationships of clinical isolates. *J Clin Microbiol* **41**, 1801-1804.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-425.

Santos, S. R. & Ochman, H. (2004). Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ Microbiol* **6**, 754-759.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol Biol Evol* **6**, 526-538.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference, *Molecular Biology and Evolution,* **16**, 1114-1116.

Skinner, D., Keffer C. (1941). Significance of bacteremia caused by *Staphylococcus aureus. Arch Intern Med* **68**, 851-875.

Smith, J. M. (1992). Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126-129.

Smith, H. O., Danner, D. B. & Deich, R. A. (1981). Genetic transformation. *Annu Rev Biochem* **50**, 41-68.

Smith, J. M., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993). How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**, 4384-4388.

Spratt, B. G. (1988). Hybrid penicillin-binding proteins in penicillin-resistant strains of *Neisseria gonorrhoeae. Nature* **332**, 173-176.

Spratt, B. G., Hanage, W. P. & Feil, E. J. (2001). The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr Opin Microbiol* **4**, 602-606.

Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S. & Musser, J. M. (1997). Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci U S A* **94**, 9869-9874.

Stachel, S. E. & Zambryski, P. C. (1986). *Agrobacterium tumefaciens* and the susceptible plant cell: a novel adaptation of extracellular recognition and DNA conjugation. *Cell* **47**, 155-157.

Stephens, J. C. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol Biol Evol* **2**, 539-556.

Stewart, G. J. & Carlson, C. A. (1986). The biology of natural transformation. *Annu Rev Microbiol* **40**, 211-235.

Suerbaum, S., Smith, J. M., Bapumia, K., Morelli, G., Smith, N. H., Kunstmann, E., Dyrek, I. & Achtman, M. (1998). Free recombination within *Helicobacter pylori*. *Proc Natl Acad Sci U S A* **95**, 12619-12624.

Sullam, P. M., Bayer, A. S., Foss, W. M. & Cheung, A. L. (1996). Diminished platelet binding in vitro by Staphylococcus aureus is associated with reduced virulence in a rabbit model of infective endocarditis. *Infect Immun* **64**, 4915-4921.

Suzuki, Y., Glazko, G. V. & Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci U S A* **99**, 16138-16143.

Swofford, D, L. (2000). PAUP and other methods. Phylogenetic analysis using parsimony, 4th ed. Sinauer Associates, Sunderland, Mass.

Teichmann, S. A. & Mitchison, G. (1999). Is there a phylogenetic signal in prokaryote proteins? *J Mol Evol* **49**, 98-107.

te Riele, H. P. & Venema, G. (1982). Molecular fate of heterologous bacterial DNA in competent *Bacillus subtilis*. I. Processing of B. pumilus and B. licheniformis DNA in B. subtilis. *Genetics* **101**, 179-188.

te Riele, H. P. & Venema, G. (1984). Molecular fate of heterologous bacterial DNA in competent *Bacillus subtilis*: further characterization of unstable association between donor and recipient DNA and the involvement of the cellular membrane. *Mol Gen Genet* **195**, 200-208.

Terry, C. E., McGinnis, L. M., Madigan, K. C., Cao, P., Cover, T. L., Liechti, G. W., Peek, R. M. Jr., Forsyth, M. H., (2005). Genomic comparisons of cag pathogenicity island (PAI)-positive and -negative *Helicobacter pylori* strains: identification of novel markers for cag PAI-positive strains. *Infect Immun* **73**, 3794-3798.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876-4882.

Trzcinski, K., Thompson, C. M. & Lipsitch, M. (2004). Single-step capsular transformation and acquisition of penicillin resistance in *Streptococcus pneumoniae*. *J Bacteriol* **186**, 3447-3452.

Tung, H., Guss, B., Hellman, U., Persson, L., Rubin, K. & Ryden, C. (2000). A bone sialoprotein-binding protein from *Staphylococcus aureus*: a member of the staphylococcal Sdr family. *Biochem J* **345** Pt 3, 611-619.

Ueda, K., Seki, T., Kudo, T., Yoshida, T. & Kataoka, M. (1999). Two distinct mechanisms cause heterogeneity of 16S rRNA. *J Bacteriol* **181**, 78-82.

Uhlen, M., Guss, B., Nilsson, B., Gotz, F. & Lindberg, M. (1984). Expression of the gene encoding protein A in Staphylococcus aureus and coagulase-negative staphylococci. *J Bacteriol* **159**, 713-719.

Vulic, M., Dionisio, F., Taddei, F. & Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci U S A* **94**, 9763-9767.

Wang, G. C. & Wang, Y. (1997). Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl Environ Microbiol* **63**, 4645-4650

Wann, E. R., Gurusiddappa, S. & Hook, M. (2000). The fibronectin-binding MSCRAMM FnbpA of *Staphylococcus aureus* is a bifunctional protein that also binds to fibrinogen. *J Biol Chem* **275**, 13863-13871.

Ward, P. D. & Turner, W. H. (1980). Identification of staphylococcal Panton-Valentine leukocidin as a potent dermonecrotic toxin. *Infect Immun* **28**, 393-397.

Weidenmaier, C., Kokai-Kun, J. F., Kristian, S. A. & other authors (2004). Role of teichoic acids in *Staphylococcus aureus* nasal colonization, a major risk factor in nosocomial infections. *Nat Med* **10**, 243-245

Weinstein, H. J. (1959). The relation between the nasal-staphylococcal-carrier state and the incidence of postoperative complications. *N Engl J Med* **260**, 1303-1308.

Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* **51**, 221-271.

Yap, W. H., Zhang, Z. & Wang, Y. (1999). Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol* **181**, 5201-5209.

Yu, V. L., Goetz, A., Wagener, M., Smith, P. B., Rihs, J. D., Hanchett, J. & Zuravleff, J. J. (1986). *Staphylococcus aureus* nasal carriage and infection in patients on hemodialysis. Efficacy of antibiotic prophylaxis. *N Engl J Med* **315**, 91-96.

Zawadzki, P., Roberts, M. S. & Cohan, F. M. (1995). The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. *Genetics* **140**, 917-932.

APPENDIX A

A1. PCR and sequencing primers for Chapter 3

Category	gene	forward primer 5' - 3'	reverse primer 5' - 3'	sequence length
CELLULAR ENVELOPE	*vicK*	gtg tga aga ttt atg gta atg atg	cct cat ctt tcc tag cac ctt c	384
	SA0272	gcg aca ggt cga gca tta gg	acc gat aga ttt tac ctg tc	450
	SA0817	cct gat gga ga(ct) gt(ct) gt	cc(at) ata cc(ct) ata ag(ag) ggc ata cg	489
	pbp2	gac caa caa gtt ggt gc	gcg tt(ag) tta taa gta cca cc	474
	AapA	tgt tgg att att ac(at) ggt atg gc	cat ga(at) ga (ag) gca gca gc	423
HOUSEKEEPING	SA0008	gaa cag att aat gtg gca gcg tat g	ctg tag ttc cca tag ata cgt gat c	429
	SA0143	gaa cgg aac agc ctc aaa tta aaa	aat cgc ttg tag act caa acc tc	432
	SA0224	gag aag gat caa gaa cga ttc ttt gg	atg ttc ttt atc ctt ttt gta aaa tcc	456
	yqiL	cag cat aca gga cac cta ttg gc	cgt tga gga atc gat act gga ac	516
	pta	gtt aaa atc gta tta cct gaa gg	gac cct ttt gtt gaa aaa agc tta a	474
	tpi	tcg ttc att ctg aac gtc gtg aa	ttt gca cct tct aac aat tgt ac	402
	gmk	atc gtt tta tcg gga cca tc	tca tta act aca acg taa tcg ta	429
	gtpF	cta gga act gca atc tta atc c	tgg taa aat cgc atg tcc aat tc	465
	aroE	atc gga aat cct att tca cat tc	ggt gtt gta tta ata acg ata tc	456
	hemH	a) agg aag gta tta tca atg act a	a) aac tat gtg ccg aaa caa cta a	819
	-	b) tca gta gga tca tat gac aaa cg	b) gtg tat tcg gca ttt ttg gac g	
	leuB	a) gcc cta cct ggt gat gga atc gg	a) cgc cac aat tta cta gaa gc	849
	-	b) gtc cgt gaa ttg aca agt gg	b) ccg cct aaa tct gct gtc g	
	hutI	a) cga gaa cat gag atg tca tta a	a) caa ttg cta att cca gtc cgc c	807
	-	b) aaa gag gca agt tca aat gag gc	b) cac ctg tgt taa ttg tac ccg c	
	arcC	ttg att cac cag cgc gta ttg tc	agg tat ctg ctt caa tca gcg	456
INFORMATION PATHWAYS	*sarA*	atg gca att aca aaa atc aat gat	tag ttc aat ttc gtt gtt tgc ttc	294
	serS	ggt gac gat att aaa gaa aaa gat	cct ctt gta tct cta cc	453
	dnaC	gcg gta ggt aaa tta tc	att ata ttg ttt cat aaa gtg t	414
	SA0189	ggt ttt gat agt aaa gta ctg	ttc aca tgg aag tgt gat aaa	432
	tufA	gta tct gct gct gac ggt cca atg	gct aat act tga cca cgt tgt ac	462
	agrC	tga aat gcg (ct)aa gtt ccg (at)ca tga	ctt tta aag ttg ata (ag)ac cta aac c	390
	sigB	gcg aaa gag tcg aaa tca gc	agc gta aca gtt gaa cca tc	441
	luxS	atg ac(at) aaa atg aat gt(at) gaa ag	ttt tcc tgt acc gaa aac atc	384
ORPHANS	SA0268	gga tta aac cgt tca ggt gca ttt aa	cca ata cag cta aat taa tta tc	474
	SA0740	atg att aat at(ct) att tca gct ata gg	cgg tcg att tga cct ttt a	456
	SA1619	gct atc gtt gca atc aca tta tc	aca cga ttt tta tcg ttt tta tc	417
	SA1621	gag aaa aac gaa tat aca gc	cta gac tca agg aaa tca tat at	456
	SA2445	gag gtt att tat cag cga tac g	ggt atg tca ttt tga tgg cgc	459
	SA0139	gta aga gaa gat aat gga g	ggt ggg tta taa tcg tta tc	426
UNKNOWN FUNCTION	SA0778	gca cct gat gtt ggg gat tat aaa	cca cc(at) acc ata ctg c	456
	SA0013	ttt aaa agg tat agt tcc gat cat tcg	caa ttt cgt tca tca cgc gtc g	435
	SA0100	gtg ggt cag att gtc tta tca atg	ccg ata cgt tca aga aca cc	444
	SA0775	gg(ac) ttt aca ggt cat at(ac) att cg	cgt act act gga tct aag aaa cc	405
	SA0275	gag aaa caa tca ttt gat gct tag ttg	ctt cag aac gtt tat cat tc	450
	SA1544	gtt gaa gtt aaa gca gtc aat att g	cct ttt aag atg gcc ttg acc acc agc	492/495
	SA2439	gat gaa atc ata aaa cga gct aa	ctt ttt ata cgc tct ttc gtc tta	410/422/434
OTHER	SA0117	ggt tag ggg aaa gtg atg	cca acc att ttc atg aat gag cat c	438
	16SrDNA	ggc agc agt agg gaa tct tcc gc	ccc gtc aat tcc ttt gag ttt caa cc	470

APPENDIX A

A2. d_S/d_N for GROUP1 strains

Category	gene	d_S	d_N	d_S/d_N
CELLULAR ENVELOPE	*vicK*	0.037	0.000	α
	SA0272	0.080	0.011	7.30
	SA0817	0.044	0.012	3.70
	pbp2	0.021	0.003	7.00
	AatpA	0.056	0.001	56.00
HOUSEKEEPING	SA0008	0.017	0.002	8.50
	SA0143	0.009	0.000	α
	SA0224	0.033	0.003	11.00
	yqiL	0.012	0.001	12.00
	pta	0.024	0.003	8.00
	tpi	0.016	0.004	4.00
	gmk	0.015	0.000	α
	glpF	0.020	0.002	10.00
	aroE	0.024	0.002	12.00
	hemH	0.033	0.002	16.50
	leuB	0.036	0.005	7.20
	hutI	0.041	0.003	13.70
	arcC	0.021	0.004	5.30
INFORMATION PATHWAYS	*serS*	0.017	0.000	α
	dnaC	0.039	0.000	α
	SA0189	0.023	0.001	23.00
	tufA	0.006	0.001	6.00
	agrC	0.215	0.015	14.30
	sigB	0.004	0.000	α
	luxS	0.028	0.002	14.00
ORPHANS	SA0268	0.000	0.009	α
	SA0740	0.027	0.006	4.50
	SA1619	0.051	0.023	2.20
	SA1621	0.036	0.009	4.00
	SA2445	0.046	0.009	5.10
	SA0139	0.037	0.008	4.60
UNKNOWN FUNCTION	SA0778	0.012	0.001	12.00
	SA0013	0.040	0.002	20.00
	SA0100	0.036	0.000	α
	SA0775	0.020	0.000	α
	SA0275	0.113	0.005	22.60
	SA1544	0.030	0.002	13.90
	SA2439	0.009	0.026	0.35

APPENDIX A

A3. d_S/d_N for GROUP 2 strains only

Category	gene	d_S	d_N	d_S/d_N
CELLULAR ENVELOPE	*vicK*	0.029	0	α
	SA0272	0.036	0.002	18.00
	SA0817	0.036	0.002	18.00
	pbp2	0.015	0.000	α
	AapA	0.017	0.000	α
HOUSEKEEPING	SA0008	0.01	0.001	10.00
	SA0143	0.014	0.000	α
	SA0224	0.024	0.006	4.00
	yqiL	0.017	0.003	5.67
	pta	0.009	0.003	3.00
	tpi	0.021	0.002	10.50
	gmk	0.015	0.003	5.00
	glpF	0.003	0.000	α
	aroE	0.008	0.006	1.33
	hemH	0.016	0.001	16.00
	leuB	0.009	0.001	9.00
	hutI	0.058	0.001	58.00
	arcC	0.016	0.003	5.33
INFORMATION PATHWAYS	*serS*	0.011	0.000	α
	dnaC	0.032	0.000	α
	SA0189	0.015	0.002	7.50
	tufA	0.001	0.001	1.00
	agrC	0.108	0.011	9.82
	sigB	0.005	0.001	5.00
	luxS	0.013	0.000	α
ORPHANS	SA0268	0.006	0.003	2.00
	SA0740	0.01	0.006	1.67
	SA1619	0.068	0.031	2.19
	SA1621	0.072	0.018	4.00
	SA2445	0.023	0.006	3.83
	SA0139	0.025	0.012	2.08
UNKNOWN FUNCTION	SA0778	0.004	0.000	α
	SA0013	0.039	0.000	α
	SA0100	0.038	0.000	α
	SA0775	0.012	0.010	1.20
	SA0275	0.074	0.002	37.00
	SA1544	0.004	0.002	2.00
	SA2439	0.014	0.008	1.75

APPENDIX A

A4. d_S/d_N for GROUP 3 strains only

Category	gene	d_S	d_N	d_S/d_N
CELLULAR ENVELOPE	*vicK*	0.021	0	α
	SA0272	0.041	0.007	5.86
	SA0817	0.024	0.030	0.80
	pbp2	0.015	0.000	α
	AatpA	0.010	0.000	α
HOUSEKEEPING	SA0008	0.009	0.001	9.00
	SA0143	0.011	0.004	2.75
	SA0224	0.024	0.001	24.00
	yqiL	0.016	0.001	16.00
	pta	0.008	0.000	α
	tpi	0.027	0.007	3.86
	gmk	0.000	0.000	α
	glpF	0.014	0.000	α
	aroE	0.036	0.008	4.50
	hemH	0.023	0.002	11.50
	leuB	0.007	0.001	7.00
	hutI	0.013	0.002	6.50
	arcC	0.016	0.003	5.33
INFORMATION PATHWAYS	*serS*	0.004	0.000	α
	dnaC	0.035	0.001	35.00
	SA0189	0.022	0.003	7.33
	tufA	0.012	0.000	α
	agrC	0.010	0.000	α
	sigB	0.009	0.000	α
	luxS	0.005	0.002	2.50
ORPHANS	SA0268	0.000	0.008	0.00
	SA0740	0.033	0.004	8.25
	SA1619	0.033	0.011	3.00
	SA1621	0.033	0.008	4.13
	SA2445	0.006	0.009	0.67
	SA0139	0.014	0.017	0.82
UNKNOWN FUNCTION	SA0778	0.000	0.000	α
	SA0013	0.000	0.001	α
	SA0100	0.113	0.005	22.60
	SA0775	0.015	0.000	α
	SA0275	0.033	0.001	33.00
	SA1544	0.007	0.004	1.93
	SA2439	0.018	0.014	1.29

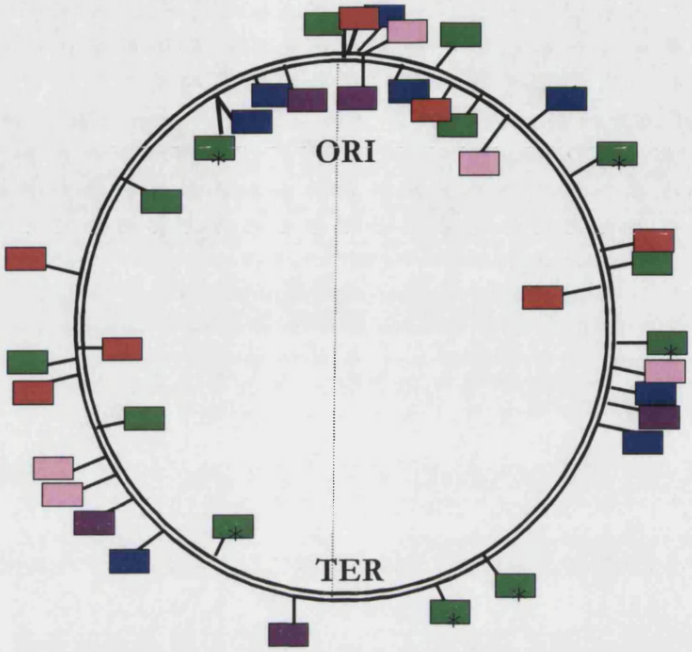
APPENDIX A

A5. Population-scaled recombination rate (ϱ) for 3 population groups

Category	gene	GROUP 1	GROUP 2	GROUP 3
CELLULAR ENVELOPE	*vicK*	6.06	12.12	16.16
	SA0272	2.02	0.00	3.03
	SA0817	4.04	4.04	3.03
	pbp2	30.30	15.15	0.00
	AapA	0.00	16.16	8.08
HOUSEKEEPING	SA0008	2.02	2.02	0.00
	SA0143	2.02	2.02	0.00
	SA0224	3.03	4.04	0.00
	yqiL	24.24	3.03	4.04
	pta	3.03	0.00	6.06
	tpi	5.05	6.06	8.08
	gmk	4.04	6.06	-
	glpF	2.02	2.02	0.00
	aroE	0.00	8.08	2.02
	hemH	7.70	6.06	5.05
	leuB	0.00	9.09	0.00
	hutI	0.00	0.00	0.00
	arcC	6.06	17.17	2.02
INFORMATION PATHWAYS	*serS*	0.00	16.16	-
	dnaC	20.20	18.18	8.08
	SA0189	6.06	48.49	13.13
	tufA	0.00	21.12	4.04
	agrC	3.03	0.00	70.71
	sigB	-	2.02	0.00
	luxS	3.03	0.00	0.00
ORPHANS	SA0268	0.00	0.00	2.02
	SA0740	1.01	4.04	19.19
	SA1619	4.04	16.16	13.13
	SA1621	2.02	10.10	0.00
	SA2445	3.03	0.00	7.07
	SA0139	0.00	16.16	0.00
UNKNOWN FUNCTION	SA0778	100.00	9.09	0.00
	SA0013	18.18	6.06	8.08
	SA0100	0.00	7.07	6.06
	SA0775	0.00	2.02	0.00
	SA0275	0.00	9.09	3.03
	SA1544	2.02	3.03	0.00
	SA2439	8.08	7.07	1.01

APPENDIX A

A6. Genomic representation of locations of selected genes



Housekeeping
Cellular envelope/processes
Information Pathways
ORPHANS
Unknown Function
* MLST loci

APPENDIX B

B1. PCR primers and conditions

Allele	forward primer 5' - 3'	reverse primer 5' - 3'	sequence length
sdrE	cag taa atg tgt caa aag a	ttg act acc agc tat atc	767 bp
bbp	cag taa atg tgt caa aag a	tac acc ctg ttg aac tg	1055 bp

Gene	forward primer 5' - 3'	reverse primer 5' - 3'	sequence length
sdrD	gga aat aaa gtt gaa gtt tc	act ttt gtc atc aac tgt aat	500 bp

PCR conditions

1. 94°C for 3.00 mins
2. 94°C for 1.00 min
3. 45°C for 1.00 min
4. 72°C for 1.00 min
5. Go to step 2, 34 cycles
6. 72°C for 10.00 mins
7. 4°C forever

APPENDIX B

B2. Data for *sdrE* and *bbp* presence testing of Oxford Collection isolates

ORIGIN	POP'N GRP	STRAIN	ST	*sdrE* allele	*bbp* allele	clonal complex
D	1	210	21	1	0	22
D	1	3	22	0	0	22
C	1	13	22	1	0	22
H	1	65	22	1	0	22
D	1	75	22	0	0	22
D	1	126	22	1	0	22
D	1	140	22	1	0	22
D	1	141	22	1	0	22
D	1	182	22	1	0	22
HMRSA	1	182	22	0	0	22
D	1	187	22	1	0	22
HMRSA	1	208	22	1	0	22
D	1	375	22	1	0	22
H	1	383	22	1	0	22
C	1	414	22	0	0	22
D	1	437	22	1	0	22
D	1	514	22	1	0	22
D	1	548	22	1	0	22
C	1	640	22	1	0	22
CMRSA	1	720	22	1	0	22
HMRSA	1	724	22	1	0	22
C	1	49	23	1	0	22
D	1	118	44	1	0	22
H	1	40	60	1	0	22
D	1	308	61	1	0	22
D	1	374	134	1	0	22
D	1	401	2	0	1	30
D	1	440	24	0	1	30
D	1	33	30	0	1	30
H	1	73	30	0	0	30
C	1	101	30	0	1	30
D	1	107	30	0	1	30
D	1	109	30	0	1	30
D	1	117	30	0	1	30
D	1	121	30	0	1	30
H	1	133	30	0	1	30
H	1	144	30	0	1	30
D	1	150	30	0	1	30
D	1	194	30	0	1	30
C	1	209	30	0	1	30
D	1	214	30	0	1	30
C	1	215	30	0	1	30
D	1	217	30	0	1	30
D	1	257	30	1	1	30
D	1	275	30	0	1	30
D	1	277	30	0	1	30
C	1	279	30	0	1	30
H	1	307	30	1	1	30
D	1	324	30	0	1	30
D	1	328	30	0	1	30
D	1	332	30	0	1	30
D	1	339	30	0	1	30
D	1	363	30	0	1	30
D	1	369	30	0	1	30
D	1	373	30	0	1	30
D	1	384	30	0	1	30

APPENDIX B

ORIGIN	POP'N GRP	STRAIN	ST	*sdrE* allele	*bbp* allele	clonal complex
D	1	388	30	0	1	30
C	1	396	30	0	1	30
H	1	406	30	0	1	30
D	1	412	30	0	1	30
H	1	415	30	0	1	30
D	1	438	30	0	1	30
D	1	477	30	0	1	30
D	1	484	30	0	0	30
D	1	495	30	0	1	30
D	1	502	30	0	1	30
D	1	504	30	0	1	30
C	1	507	30	0	1	30
H	1	509	30	0	1	30
H	1	513	30	0	1	30
D	1	528	30	0	1	30
D	1	557	30	0	1	30
D	1	560	30	0	1	30
C	1	736	30	0	1	30
C	1	739	30	0	1	30
H	1	908	30	0	1	30
C	1	390	31	0	1	30
H	1	140	32	0	1	30
H	1	399	33	0	1	30
D	1	90	34	0	1	30
D	1	153	34	0	1	30
C	1	160	34	0	1	30
D	1	233	34	0	0	30
D	1	289	34	0	1	30
D	1	330	34	0	1	30
D	1	350	34	0	1	30
D	1	476	34	0	1	30
D	1	501	34	0	1	30
D	1	555	34	0	1	30
D	1	564	34	0	1	30
C	1	959	34	0	0	30
HMRSA	1	21	36	0	1	30
HMRSA	1	41	36	0	1	30
HMRSA	1	45	36	0	1	30
HMRSA	1	69	36	1	1	30
HMRSA	1	119	36	0	1	30
HMRSA	1	123	36	0	1	30
HMRSA	1	252	36	0	1	30
HMRSA	1	281	36	0	1	30
HMRSA	1	312	36	0	1	30
HMRSA	1	325	36	1	1	30
HMRSA	1	354	36	0	1	30
HMRSA	1	370	36	0	1	30
HMRSA	1	378	36	0	1	30
HMRSA	1	393	36	0	1	30
HMRSA	1	448	36	0	1	30
HMRSA	1	455	36	1	1	30
HMRSA	1	504	36	0	1	30
HMRSA	1	721	36	0	1	30
HMRSA	1	831	36	1	1	30
HMRSA	1	858	36	0	1	30
HMRSA	1	906	36	1	1	30

APPENDIX B

ORIGIN	POP'N GRP	STRAIN	ST	*sdrE* allele	*bbp* allele	clonal complex
C	1	563	37	0	1	30
HMRSA	1	137	38	0	1	30
H	1	39	39	0	1	30
D	1	42	39	0	0	30
D	1	74	39	0	1	30
D	1	106	39	0	1	30
C	1	114	39	0	1	30
H	1	304	39	0	1	30
D	1	315	39	0	1	30
D	1	349	39	0	1	30
C	1	364	39	0	1	30
D	1	432	39	1	1	30
D	1	453	39	0	0	30
D	1	455	39	0	0	30
D	1	494	39	0	1	30
D	1	546	39	0	1	30
D	1	565	39	0	0	30
H	1	599	39	0	1	30
C	1	767	39	1	1	30
H	1	888	39	0	1	30
H	1	952	39	0	1	30
C	1	253	40	0	1	30
H	1	301	41	0	1	30
C	1	427	42	0	1	30
D	1	354	43	0	1	30
D	1	318	57	0	1	30
D	1	224	77	0	0	30
C	1	95	4	1	0	45
D	1	219	45	1	0	45
C	1	233	45	1	0	45
H	1	295	45	1	0	45
D	1	304	45	1	0	45
H	1	321	45	1	0	45
D	1	334	45	0	0	45
D	1	366	45	1	0	45
D	1	368	45	1	0	45
D	1	395	45	1	0	45
H	1	456	45	1	0	45
H	1	617	45	1	0	45
C	1	730	45	1	0	45
C	1	900	45	1	0	45
D	1	164	46	1	0	45
D	1	170	46	1	0	45
H	1	486	46	0	0	45
C	1	96	47	0	0	45
D	1	283	47	1	0	45
D	1	284	47	1	0	45
D	1	386	47	1	0	45
H	1	442	47	1	0	45
C	1	474	47	1	0	45
H	1	481	47	1	0	45
D	1	513	47	1	0	45
H	1	624	48	0	0	45
D	1	49	53	0	0	45
D	1	98	54	1	0	45
D	1	127	54	1	0	45
D	1	451	54	1	0	45

APPENDIX B

ORIGIN	POP'N GRP	STRAIN	ST	*sdrE* allele	*bbp* allele	clonal complex
H	1	19	10	0	0	m
D	1	139	145	1	0	m
D	1	470	207	1	0	s
D	2	5	1	0	0	1
C	2	38	1	1	0	1
C	2	71	1	1	0	1
C	2	98	1	1	0	1
C	2	162	1	1	0	1
D	2	285	1	1	0	1
D	2	325	1	1	0	1
H	2	462	1	1	0	1
C	2	476	1	1	0	1
H	2	512	1	1	0	1
C	2	1472	1	1	0	1
H	2	148	3	1	0	1
D	2	473	69	1	0	1
H	2	52	188	1	0	1
D	2	467	188	1	0	1
D	2	552	188	1	0	1
D	2	10	5	1	0	5
D	2	52	5	1	0	5
D	2	80	5	1	0	5
D	2	102	5	1	0	5
D	2	128	5	1	0	5
H	2	157	5	1	0	5
D	2	174	5	1	0	5
D	2	181	5	1	0	5
D	2	237	5	1	0	5
D	2	243	5	1	0	5
D	2	409	5	1	0	5
C	2	433	5	1	0	5
D	2	441	5	1	0	5
H	2	451	5	1	0	5
H	2	466	5	1	0	5
D	2	483	5	1	0	5
C	2	511	5	1	0	5
C	2	521	5	1	0	5
H	2	811	5	1	0	5
C	2	56	6	1	0	5
C	2	155	6	1	0	5
D	2	316	11	1	0	5
D	2	21	8	1	0	8
H	2	67	8	1	0	8
D	2	83	8	1	0	8
C	2	125	8	1	0	8
D	2	137	8	1	0	8
D	2	268	8	1	0	8
H	2	315	8	1	0	8
H	2	326	8	1	0	8
D	2	391	8	1	0	8
D	2	414	8	1	O	8
D	2	419	8	1	0	8
C	2	434	8	1	0	8
D	2	521	8	0	0	8
H	2	591	8	1	0	8
C	2	827	8	1	0	8

APPENDIX B

ORIGIN	POP'N GRP	STRAIN	ST	*sdrE* allele	*bbp* allele	clonal complex
C	2	850	8	1	0	8
H	2	116	9	1	0	9
H	2	150	9	1	0	9
H	2	169	9	1	0	9
D	2	295	9	1	0	9
D	2	410	27	1	0	9
D	2	472	109	0	0	9
C	2	154	14	1	0	15
D	2	16	15	1	0	15
D	2	31	15	1	0	15
D	2	61	15	1	0	15
D	2	77	15	1	0	15
D	2	95	15	1	0	15
D	2	143	15	1	0	15
D	2	144	15	1	0	15
D	2	158	15	1	0	15
D	2	197	15	1	0	15
C	2	207	15	1	0	15
D	2	307	15	1	0	15
D	2	341	15	1	0	15
C	2	357	15	1	0	15
H	2	382	15	1	0	15
H	2	458	15	1	0	15
D	2	462	15	1	0	15
D	2	469	15	1	0	15
D	2	478	15	1	0	15
D	2	527	15	1	0	15
C	2	686	15	1	0	15
H	2	783	15	1	0	15
H	2	410	16	1	0	15
H	2	291	18	1	0	15
H	2	964	18	1	0	15
D	2	336	35	1	0	15
D	2	78	56	1	0	15
D	2	508	58	1	0	15
D	2	27	169	1	0	15
D	2	340	178	1	0	15
C	2	16	25	1	0	25
C	2	25	25	1	0	25
D	2	56	25	1	0	25
D	2	57	25	0	0	25
D	2	111	25	1	0	25
H	2	159	25	1	0	25
C	2	197	25	1	0	25
H	2	205	25	1	0	25
D	2	279	25	1	0	25
C	2	283	25	1	0	25
D	2	306	25	1	0	25
H	2	309	25	1	0	25
D	2	337	25	1	0	25
H	2	352	25	1	0	25
C	2	437	25	1	0	25
C	2	449	25	1	0	25
D	2	489	25	1	0	25
D	2	559	25	1	0	25
C	2	593	25	1	0	25

APPENDIX B

ORIGIN	POP'N GRP	STRAIN	ST	*sdrE* allele	*bbp* allele	clonal complex
H	2	597	25	1	0	25
C	2	701	25	1	0	25
H	2	112	26	1	0	25
H	2	118	28	1	0	25
H	2	117	12	1	0	m
C	2	126	12	1	0	m
D	2	244	12	0	0	m
H	2	303	12	0	0	m
D	2	329	12	0	0	m
D	2	342	12	0	0	m
D	2	446	12	0	0	m
H	2	525	12	0	0	m
C	2	837	12	0	0	m
H	2	402	13	0	0	m
C	2	2	7	0	1	s
D	2	7	20	1	0	s
D	2	17	20	1	0	s
H	2	42	20	1	0	s
D	2	93	20	1	0	s
H	2	863	20	1	0	s
D	2	97	55	1	0	s
D	2	302	97	1	0	s
D	2	547	97	1	0	s
D	2	215	101	0	1	s
D	2	346	101	0	1	s
D	2	358	101	0	1	s
D	2	456	101	0	1	s
D	3	371	29	0	1	121
C	3	3	51	1	0	121
D	3	199	121	0	1	121
D	3	365	121	0	1	121
D	3	422	121	0	1	121
H	3	560	121	0	1	121
D	3	566	121	0	1	121
D	3	54	123	0	1	121
D	3	115	123	0	1	121
D	3	291	123	0	1	121
C	3	316	49	1	0	s
H	3	707	49	1	0	s
D	3	274	17	1	0	s
H	3	417	50	1	0	s
D	3	535	59	1	0	s
D	3	551	59	1	0	s
D	3	22	182	0	0	s

Strains which do not belong to a clonal complex:

m – minor group
s - singletons

APPENDIX B

B3. Data for *sdrE* and *bbp* presence testing of Nottingham Collection isolates

	STRAIN	*sdrE* allele	*bbp* allele
B	1	1	0
B	4	1	0
B	7	1	0
B	14	0	0
B	17	0	0
B	20	0	0
B	22	1	0
B	26	1	0
B	28	1	0
B	35	0	1
B	38	0	0
B	41	0	0
B	44	0	1
B	62	1	0
B	67	1	0
B	69	1	0
B	71	1	0
B	72	1	0
B	74	1	0
B	78	0	0
B	87	1	0
B	106	1	0
B	109	0	0
B	111	1	0
B	113	0	0
B	116	0	0
B	117	1	0
B	119	0	0
B	126	0	0
B	130	1	0
B	139	0	0
B	169	1	0
B	171	1	0
B	173	1	0
B	176	1	0
B	177	0	1
B	178	0	1
B	179	0	1
B	180	1	0
B	202	1	0
B	206	1	0
B	210	0	1
B	215	0	0
B	216	0	0
B	217	0	0
B	221	1	0
B	222	0	0
B	223	1	0
B	226	0	1
B	240	1	0
B	241	0	1
B	248	0	0
B	271	0	0
B	273	0	0
B	279	1	0
B	281	0	0

APPENDIX B

	STRAIN	*sdrE* allele	*bbp* allele
B	283	1	0
B	284	0	1
B	285	0	1
B	288	0	1
B	293	1	0
B	294	1	0
B	306	1	0
B	309	0	0
B	310	1	0
B	313	0	1
B	315	0	0
B	318	1	0
B	342	0	0
B	343	1	0
B	346	1	0
B	353	1	0
B	354	0	0
B	355	0	0
B	366	0	1
B	367	0	0
B	371	1	0
B	372	0	1
B	379	0	1
B	387	1	0
B	388	1	0
B	398	1	0
B	403	1	0
B	407	0	1
B	419	0	1
B	420	0	1
B	426	1	0
B	427	1	0
B	428	1	0
B	433	1	0
B	443	0	0
B	444	0	0
B	446	1	0
B	448	0	1
B	449	1	0
B	452	0	1
B	453	1	0
B	454	0	0
B	459	0	1
B	462	1	0
B	464	1	0
B	465	1	0
B	470	1	0
B	483	1	0
B	484	0	0
B	486	1	0
B	488	0	0
B	491	1	0
B	493	1	0
B	496	1	0
B	498	0	0
B	500	0	1

APPENDIX B

	STRAIN	*sdrE* allele	*bbp* allele
B	501	1	0
B	503	1	0
B	504	0	1
B	513	0	1
B	519	0	1
B	520	0	1
B	521	1	0
B	525	0	1
B	537	1	0
B	538	1	0
B	539	0	1
B	541	1	0
B	552	1	0
B	558	0	1
B	563	0	1
B	565	1	0
B	567	1	0
B	568	1	0
B	574	0	0
B	575	1	0
B	577	1	0
B	621	1	0
B	625	1	0
B	632	1	0
B	634	0	1
B	638	1	0
B	643	0	1
B	645	0	1
B	646	1	0
B	650	1	0
B	654	1	0
B	664	1	0
B	666	1	0
B	670	1	0
B	671	0	0
B	681	0	1
B	689	1	0
B	691	1	0
B	701	1	0
B	702	1	0
B	703	0	1
B	705	0	1
B	710	0	0
B	712	0	0
B	725	0	1
B	728	0	1
B	731	0	0
B	737	1	0

APPENDIX B

B4. PCR testing of *sdrE* and *bbp* orientations in isolates with both

Target allele	Primer 5' - 3'	tests arrangement
sdrE	ggt agt gaa aat aac gg	1, 2
sdrE	gca gct tta gct tct tgg ttc cc	3, 4
bbp	gca ttg aca ttc tca tat cta tc	1, 3
bbp	cta tgt atg gga aga tac g	2, 4

PCR conditions

1. 94°C for 3.00 mins
2. 94°C for 1.00 min
3. 47°C for 1.00 min
4. 72°C for 1.00 min
5. go to step 2, 34 cycles
6. 72°C for 10.00 mins
7. 4°C forever

APPENDIX B

B5. Data for *sdrD* (and *sdrE* locus alleles) using diverse strains from the Oxford collection of isolates

ORIGIN	POP'N GRP	STRAIN	ST	*sdrD* locus	*sdrE* allele	*bbp* allele	clonal complex
C	1	640	22	1	1	0	22
CMRSA	1	720	22	1	1	0	22
C	1	101	30	0	0	1	30
HMRSA	1	325	36	0	1	1	30
HMRSA	1	831	36	0	1	1	30
H	1	295	45	0	1	0	45
H	1	19	10	0	0	0	m
D	1	470	207	0	1	0	s
H	2	512	1	1	1	0	1
H	2	466	5	1	1	0	5
H	2	591	8	1	1	0	8
H	2	116	9	1	1	0	9
H	2	783	15	1	1	0	15
C	2	437	25	1	1	0	25
H	2	402	13	1	0	0	m
D	2	17	20	1	1	0	s
D	2	97	55	1	1	0	s
D	2	547	97	1	1	0	s
D	2	456	101	1	0	1	s
D	3	365	121	0	0	1	121
H	3	560	121	0	0	1	121
H	3	707	49	0	1	0	s
D	3	274	17	0	1	0	s
H	3	417	50	0	1	0	s
D	3	535	59	0	1	0	s
EMRSA	2	3	5	1	1	0	5
EMRSA	2	4	239	1	1	0	8
EMRSA	2	9	240	1	1	0	8

Strains which do not belong to a clonal complex:

m – minor group
s - singleton

APPENDIX C

C1. PCR and sequencing primers for *sdrE* and *bbp* alleles

Name	primer sequence 5' - 3'	prime direction 5' - 3'
ORFX	caa tta cga gca aga tta ttt gtc ga	forward
sdrEstart	ccg gga tcc tga tta aca ggc ata aaa ag	forward
sdrE1	gca aat att gat att tta aa	forward
sdrE2	gaa tca gta ttt gtt tct ttc tt	reverse
sdrE3	gta gac aat caa gtt aca gat gc	forward
sdrE4	gaa gca act gct gct ggt tgt gc	reverse
sdrE5*	act aag caa atc aca tat aca tt	forward
bbp5*	acg aaa aca att act tac aaa tt	forward
sdrE5.5	(ct)cg a(at) c aaa gaa at(ag) gac gac a(at)c g	reverse
sdrE6*	tat ttc tgt att ttg gtc aa	reverse
bbp6*	aat ttc cgt att act atc aa	reverse
sdrE7	cgg tac tgt taa acc tga aga aaa	forward
sdrE8	cca tac ata gtc acc aat ttt gt	reverse
sdrE9	agt ttc tcc gtc ttt caa acc acc g	reverse
sdrE10	tca aag atg tta agg tta cat ta	forward
sdrE11	aac cgt tgg cgt gta acc tgc tg	reverse
sdrE12	tag gtg att atg ttt ggt acg ac	forward
sdrE13	atc aag tgt gaa atc atc atg at	reverse
sdrE14	cga aga aga tac atc aga ca	forward
sdrE15	agg tgt atg ttt tcc tgc at	reverse
ORF522	cct aaa atg taa ttc ata tta tcg c	reverse
sdrEend	ccg gga tcc tta ttt gtt ttg ttt ttt gcg acg	reverse

* indicates that these primers are allele specific

APPENDIX D

D1. PCR and sequencing primers for the *arcC*, *crp* and *clfB* region

Name	forward primer 5' - 3'	reverse primer 5' - 3'
arcC_MLST	ttg att cac cag cgc gta ttg tc	agg tat ctg ctt caa tca gcg
arcCA	cgt ggt tat aga aaa gta gtt gc	cgg cgt tgt gtc act gtt cg
arcCB	gat gat att gat gta gc	ccc act ttc aac aaa tcg tat cg
crpA	gaa cat gaa cta aag gc	gct act ttc ttg tgg cg
crpB	cag aca gta gga tac gat caa	gaa taa tat gac cag ctg ttt
crpC	cat ttg taa tag tgt aaa aat ag	gaa ata aga att ata tca att gc
clfBA	cag aat aag tat tcg att aga cg	ccc aaa tag tat agt tgc cc
clfBB	gct(ct) caa caa atg aaa cac c	ccc aaa tag tat agt tgc cc
clfBC	gta aat gat aaa gtt acg gc	cgc tgt aaa ata atc cc
clfBD	gca aag gca cct aaa tca gg	aat tcc tgc aat tgg cg
clfBE	tca gat agc tac tat gc	cca taa cgt aca aca ttc tc